LECTRANS: A Multimodal Translation Benchmark for Academic Lectures

Frederikus Hudi[*†§], **Nina Nanakorn**[*‡], **Justin Vasselli**[†],
Yusuke Sakai[†], **Hidetaka Kamigaito**[†], **Akihiro Tamura**[‡†§], **Taro Watanabe**[†]
[†]Nara Institute of Science and Technology (NAIST)
[‡]Doshisha University
frederikus.hudi.fe7@naist.ac.jp, aktamura@mail.doshisha.ac.jp

Abstract

Existing machine translation and multimodal translation benchmarks mainly focus on text-only inputs or short examples with tightly aligned image-text pairs. Academic lectures present a different challenge: spoken explanations, presentation slides, and domain-specific terminology interact across long-form discourse. LECTRANS is constructed from 383 hours of expert-taught online lectures and provides transcripts, presentation slides, and an eight-language parallel evaluation subset covering English and seven target languages. The 162-minute translation subset spans 17 subjects and contains human-corrected English transcripts, verified slide text, and human-verified translations of both transcripts and slides. We define lecture translation as two complementary tasks: *transcript translation* for spoken lecture content and *slide translation* for visually presented text. We report text-only and multimodal baselines, showing that noisy transcripts, loosely aligned slide context, and domain-specific content remain challenging for current systems. LECTRANS provides a testbed for evaluating multimodal translation in long-form academic lectures.

1 Introduction

Advances in multimodal language modeling have made machine-translated content more accessible across languages in today's globalized world (Vo et al., 2025). Existing multimodal benchmarks, including multimodal machine translation (MMT), primarily focus on tightly aligned image-text pairs, where images provide contextual grounding for short source texts (Elliott et al., 2016). However, these settings provide limited insight into model performance in long-form scenarios where speech and visual materials convey related but non-identical information, such as lecture presentations accompanied by slides (Feng et al., 2025).



Figure 1: **Motivation for LECTRANS**. The term *convenience form* is mistranslated when treated as general language. Slide context helps recover its intended meaning. Bold text denotes domain-specific terminology, while red text denotes the mistranslations.

Academic lecture translation presents unique challenges for multimodal evaluation. Lectures convey domain knowledge through spoken explanations supported by slides that may summarize, supplement, or diverge from the speech (Lee et al., 2023). For example, Figure 1 illustrates how slide context helps recover the domain-specific meaning of *"convenience form"*, which is mistranslated without this slide context. Furthermore, lecture transcripts are often generated using automatic speech recognition (ASR) systems, introducing recognition errors and disfluencies from the original speech (Wang et al., 2024), while time-stamping and alignment errors can produce mismatched segments (Teramen et al., 2024). Existing lecture datasets (Chen et al., 2024; Wang et al., 2025a) also predominantly focus on tasks such as video understanding, ASR, text-to-speech (TTS), slide generation, or video segmentation rather than translation. Because the visual and spoken channels are related but not strictly parallel, a translation system must distinguish useful contextual information from visual content unrelated to the current utterance (Yuasa et al., 2023).

[*] These authors contributed equally.
[§] Corresponding authors.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 1872–1897
November 6-10, 2026 ©2026 Association for Computational Linguistics

To investigate this setting, we introduce LECTRANS, a benchmark and scalable data creation pipeline tailored for multimodal lecture translation. The benchmark integrates curated online lecture videos, slide materials (i.e., images and extracted text), ASR outputs, human-corrected English transcripts, and human-verified translations in seven target languages. We formulate lecture translation as two complementary tasks: **transcript translation**, which targets the lecturer's speech, and **slide translation**, which targets text extracted from presentation slides. This dual-task design enables the independent evaluation of speech-derived and visually presented content while supporting the analysis of how slide context influences transcript translation. Our contributions are as follows:

- We introduce LECTRANS[1], a multimodal benchmark for academic lecture translation constructed from 383 hours of lecture videos and an eight-language parallel evaluation subset covering transcripts and slide text.[2]
- We define a dual-task evaluation setting consisting of transcript translation and slide translation, enabling separate analysis of spoken and visually presented lecture content.
- We describe a staged data construction pipeline that combines automatic processing with targeted human verification, covering ASR, ASR alignment, slide extraction & deduplication, OCR, slide-and-transcript alignment, and translation review.
- We benchmark open and proprietary multimodal models under text-only, slide-text, and slide-image settings, analyzing the effects of transcript quality, slide context, target language, and lecture subject on translation performance.

2 Related Work

Educational Video and Lecture Understanding. Several datasets have been introduced for modeling academic and educational videos. LPMD (Lee et al., 2023) provides lecture videos with presentation slides and supports tasks such as cross-modal retrieval and multimodal educational video understanding. M3AV (Chen et al., 2024) provides audio-visual academic lectures for tasks including ASR, TTS, and slide generation. MITFLD (Wang et al., 2025a) focuses on identifying key points and segmenting lecture videos. These datasets capture interactions between speech, slides, and visual structure in educational videos, but primarily focus on lecture understanding, generation, or segmentation rather than machine translation. Wang et al. (2024) introduce a multimodal scientific ASR task in which visual information from slides is used to improve the recognition of technical terminology, which is closely related to our setting, but it addresses transcription rather than translation.

Multimodal Machine Translation (MMT) is predominantly studied in settings where images provide contextual grounding for short source texts. Early shared tasks on multimodal translation and cross-lingual image description paired images with sentence-level captions, establishing image-guided translation on Multi30K (Elliott et al., 2016) as a standard evaluation setting (Specia et al., 2016). Subsequent work has studied visual attention, image-text grounding, and alternative visual inputs, including synthetic images, to examine when visual information helps translation (Yuasa et al., 2023). Surveys of MMT similarly describe that the field is centered on short inputs with relatively close image-and-text alignment (Shen et al., 2024; Feng et al., 2025). Academic lectures pose a different problem. The source text is typically obtained from ASR and may contain recognition errors, segmentation mismatches, and spoken disfluencies. Furthermore, slide content is not strictly aligned with individual utterances: slides may summarize earlier content, introduce later content, or omit details explained in speech. This setting differs from conventional sentence-level MMT with tightly aligned image-text pairs.

Speech Translation and Simultaneous Lecture Translation. Speech translation has been studied through cascaded ASR-MT systems and end-to-end models (Sperber and Paulik, 2020), with resources such as MuST-C providing multilingual translation data from TED talks (Di Gangi et al., 2019). Early lecture translation systems addressed spontaneous speech and domain mismatch (Fügen et al., 2006). Subsequent simultaneous lecture translation systems addressed segmentation and latency, while also using slides or lecture notes for domain adaptation (Müller et al., 2016). More recent work has developed training resources and evaluation methods for simultaneous translation. Simul-MuST-C (Makinae et al., 2024) uses LLMs to generate chunk-

[1] https://github.com/naist-nlp/lectrans

[2] The dataset contains derived annotations, translations, and metadata, but does not redistribute any video content.

Dataset Name	# Hours	# Courses	Task
LPMD (Lee et al., 2023)	187	10	Information Retrieval
M^{3}AV (Chen et al., 2024)	367	5*	ASR, TTS, Slide Generation
MITFLD (Wang et al., 2025a)	559	25	Video Fragmentation
LECTRANS (ours, 2026)	**383**	**17**	**Multimodal MT** (visual-audio-text)

Table 1: Comparison of existing dataset which makes use of academic data. *M^{3}AV data source is conference venues rather than online courses.

wise translations that more closely follow source word order, while Simul-COMET (Doi et al., 2026) adapts quality evaluation to account for this ordering. LECTRANS complements these efforts by examining how transcript quality and slide context affect offline academic lecture translation.

Lecture Translation and Slide-Grounded Translation. Recent work has further explored slide-grounded and multimodal translation in lecture settings. Teramen et al. (2024) investigate English-to-Japanese lecture transcript translation using slide information as visual context, showing that slide-grounded inputs can be useful in controlled lecture translation settings. GlobalizeED (Vo et al., 2025) proposes an end-to-end lecture translation pipeline that combines ASR, machine translation, and target-language speech synthesis. BOOM (Koneru et al., 2026) similarly targets multilingual lecture access by translating lecture audio and slides into synchronized outputs. While these studies demonstrate a growing interest in multimodal lecture translation, they primarily formulate the problem either as an end-to-end system pipeline or a single translation setting. Table 1 summarizes the differences between existing resources and LECTRANS.

3 LECTRANS Benchmark

LECTRANS is a multilingual benchmark for multimodal lecture translation that integrates lecture transcripts, presentation slides, and human-verified translations across seven target languages. The benchmark defines separate transcript translation and slide translation tasks for spoken and visually presented lecture content, respectively. We translate the English sources into seven diverse target languages, for further details see Appendix A.

3.1 Source Lectures

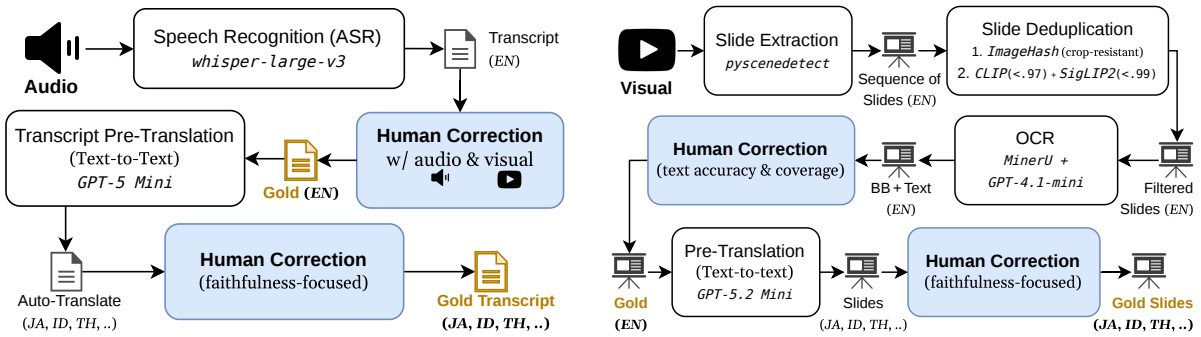
LECTRANS combines English lecture videos and slide images obtained from two sources: the Lecture Presentations Multimodal Dataset (LPMD) (Lee et al., 2023) and MIT OpenCourseWare (OCW) (Massachusetts Institute of Technology, 2001). The lectures span specialized domains and include presentation slides alongside spoken lecture content.

LPMD. We use the updated version released on October 17, 2023. This dataset includes lectures from multiple fields, including anatomy, biology, dentistry, machine learning, psychology, and speaking. The dataset provides links to YouTube channels and playlists. After preprocessing, a total of 334 lecture videos were collected. These videos are organized into 11 groups according to their source links: anatomy and psychology are each divided into two groups (anat-1, anat-2, psy-1, and psy-2), biology is divided into four groups (bio-1, bio-2, bio-3, and bio-4), while dentistry (dental), machine learning (ml-1), and speaking (speaking) consist of one group each.

MIT-OCW. We collect six undergraduate-level courses from MIT OpenCourseWare to complement the domains not represented in LPMD: Intermediate Microeconomic Theory (i-microecons) (Townsend, 2020), Introduction to CS and Programming using Python (i-cs-prog) (Bell, 2022), Introduction to Neural Computation (i-neural-comp) (Fee and Zysman, 2018), Physics in the 20th Century (i-physics) (Kaiser, 2020), Theory of Computation (theory-comp) (Sipser, 2020), and Urban Energy Systems and Policy (urban-policy) (Hsu, 2022). We include three more courses for the trainset: Introduction to Computational Thinking and Data Science (Grimson et al., 2016), Introduction to Modeling and Simulation (Buehler and Grossman, 2012), and Introduction to Psychology (Gabrieli, 2011). After preprocessing, a total of 185 lecture videos were collected.

3.2 Task Definition

Transcript Translation. Given an English lecture transcript segment, the system generates a faithful translation in the target language while preserving domain terminology, named entities, and pedagogical structure. We evaluate three input settings: (1) *No additional context*, where the model



(a) Transcript Translation task

(b) Slide Translation task

Figure 2: Translation dataset construction pipelines

receives only the transcript text; (2) *Slide text context*, where the model additionally receives text extracted from the slides displayed during the transcript segment; and (3) *Slide image context*, where the model receives the corresponding slide images as visual grounding.

Slide Translation. This task evaluates translation of visually presented instructional content at the text-region level. Each slide is processed into a set of text bounding boxes, and each text region is translated independently. We evaluate two settings: (1) *No additional context*, where translation is performed using only the extracted slide text; and (2) *Slide image context*, where the original slide image is additionally provided as visual grounding for ambiguous or incomplete text regions.

3.3 Dataset Construction Pipeline

3.3.1 Transcript Processing

Figure 2a shows the transcript processing pipeline. We obtain the audio using yt-dlp[3], and then transcribe each lecture using Whisper (Radford et al., 2022) large-v3 model. The resulting ASR segments retain timestamps from the original video, but these segments do not necessarily correspond one-to-one with the human-corrected transcript segments. Directly comparing timestamp-level ASR outputs with corrected references therefore introduces segmentation mismatches in addition to recognition errors.

To reduce discrepancies with the human-corrected transcript, we align the raw ASR text to the corrected transcript using edit distance. This alignment preserves the original ASR words while mapping them to the corrected transcript segmentation, allowing us to compare raw recognition outputs, aligned ASR, and corrected transcripts under a consistent segment structure. Additional details are provided in Appendix G.

3.3.2 Slide Processing

Figure 2b illustrates the slide processing pipeline. Lecture videos often contain long intervals with relatively static slide content, allowing representative slide frames to be extracted through scene detection and deduplication. The resulting slides serve as the primary input for slide translation and as visual context for transcript translation. Appendix E reports slide extraction statistics.

Slide Extraction. Slide images and text were obtained from lecture videos by automatically segmenting each video into scenes using the PySceneDetect library[4]. We use the preprocessing procedure of LPMD, saving the first, midpoint, and final frames of each detected scene.

Slide Deduplication. As LPMD preprocessing configuration applies finer-grained scene segmentation than PySceneDetect, the number of detected scenes exceeded the true number of lecture slides, particularly in segments containing non-slide lecture content. To remove redundant frames, we first applied a crop-resistant hash from ImageHash[5], followed by embedding-based similarity filtering using CLIP (Radford et al., 2021) and SigLIP2 (Tschannen et al., 2025).

Information Extraction. MinerU (He et al., 2024) was used to detect bounding boxes and classify slide elements into five classes: text, equations, tables, images, and others. For textual elements, the text within each bounding box was extracted and corrected by GPT-4.1 mini.

[3]https://github.com/yt-dlp/yt-dlp

[4]https://www.scenedetect.com

[5]https://github.com/JohannesBuchner/imagehash

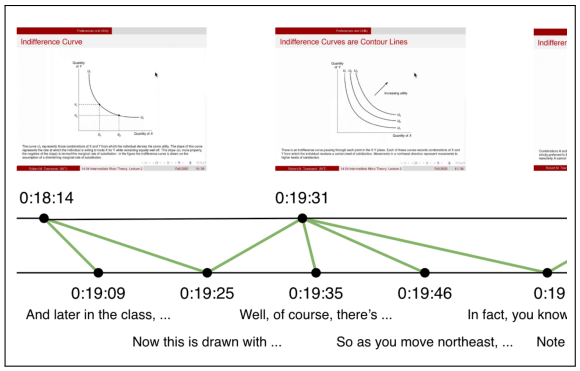


Figure 3: Illustration of Slides & Transcript alignment operation using sliding windows algorithm. Deduplicated slides and transcript segments are placed on parallel timelines, and transcript segments are linked to the slide intervals that overlap with their timestamps. This alignment provides slide text or slide images as contextual input for transcript translation as described in Section 3.3.3.

3.3.3 Slide & Transcript Alignment

After constructing the deduplicated slide sequence and extracting slide text, we align transcript segments with the slides displayed during the corresponding time intervals as shown in Figure 3. Each transcript segment retains its start and end timestamps from the ASR output, while each slide is assigned a display interval from the video timeline. We associate each transcript segment with the slide or slides whose display intervals overlap with the segment timestamp. This alignment is used to construct the multimodal input settings for transcript translation defined in Section 3.2: the *slide text context* setting uses OCR text from the aligned slides, while the *slide image context* setting uses the corresponding slide images.

3.3.4 Translation Pre-annotation

Before human verification, we generated draft translations using GPT-5 mini for transcript translation and GPT-5.2 mini for slide translation. To support evaluation across diverse language families, we include seven target languages.

3.4 Test Set: Human Annotation

3.4.1 Transcript Annotation

English Transcript Correction. Three videos from each subject were selected arbitrarily after filtering out videos with extremely high or low word counts. The ASR outputs were then reviewed and annotated by professional translators to correct transcription errors. Each transcription error was annotated with an error category (addition, omission, misspelling, or filler), and a severity level (OK, MINOR, or CRITICAL). Technical terminology was labeled as TERM. Figure 8 shows an example annotation with additional details provided in Appendix I.

Transcript Translation Correction. One video was selected from each subject, and from each selected video, a segment of approximately 10 minutes was extracted for machine translation. Professional translators who are native speakers of each language reviewed the machine-translated transcript translations to ensure faithfulness and accuracy.

3.4.2 Slide Annotation

English Slide Correction. Slide data were manually reviewed and corrected using Label Studio Community Edition (Tkachenko et al., 2020-2025). Automatically extracted slide images were reviewed by 18 annotators to remove any remaining duplicate slides and non-slide images such as blank screens or lecturer-only frames. The same annotators also verified and corrected text extracted from slide images by fixing OCR errors and adding missing text regions.

Slide Translation Correction. Finally, seven graduate students who are native speakers of the respective target languages reviewed and corrected the machine-generated slide translations.

3.5 Dataset Statistics and Splits

Table 2 summarizes the dataset statistics. The full automatically processed collection contains 23,039 minutes of English lecture video. From this collection, we construct an ASR/OCR evaluation set of 51 videos, with three videos selected from each of the 17 subject groups. For the translation evaluation subset, we further select one representative segment from each subject group, resulting in 17 translated segments. This subset contains 162 minutes of lecture content with parallel annotations in English and seven target languages.

The dataset is organized into three annotation levels. The full collection provides automatically processed lecture content at scale. The ASR/OCR evaluation set includes human-corrected English transcripts and verified slide text. The translation evaluation subset provides human-verified multilingual references for transcript and slide translation.

#	All (Automated)	Testset - ASR & OCR	Testset - Translation
Courses	20	17	17
Videos	519	51 (3×17 courses)	17 segments (1 ea.)
Minutes	23,039	2,510	162 (of 820)
Languages	EN	EN	EN,JA,ID,TH,ZH,ES,FR,DE
Words (EN)	approx. 3.7M	445,596	27,315
Unique Slides	approx. 15K	3,022	164

Table 2: Statistical overview of LECTRANS benchmark of academic lecture translation.

4 Experimental Setup

4.1 Vision-Language Models

We evaluate ten Vision-Language Models (VLM) for multilingual text-visual generation to compare performance on lecture translation tasks across multimodal architectures. Open-weight models provide reproducible baselines, while proprietary models serve as strong references.

The open-source models we evaluated include Llama 4 Scout (Meta, 2025), Llama 3.2 11B (Grattafiori et al., 2024), Gemma 3 12B (Team et al., 2025), Qwen3.5 9B (Qwen Team, 2026), Aya Vision 8B (Dang et al., 2024), InternVL3.5 8B (Wang et al., 2025b), and GLM-4.6V Flash (Team et al., 2026). The proprietary models we evaluated include GPT-4.1, GPT-5 mini, and o4-mini. Appendix B contains the full list of evaluated models.

4.2 Evaluation Metrics

We evaluate translation quality using both reference-based and reference-free metrics. Reference-based evaluation includes ChrF (Popović, 2015), SacreBLEU (Post, 2018), COMET (Rei et al., 2022a), and NIST (Doddington, 2002). NIST is included to better capture technical terminology and domain-specific expressions in lecture content. Note that NIST is included as a complementary lexical metric, and it is not a direct measure of terminology translation accuracy or consistency. For SacreBLEU evaluation on non-Latin scripts, language-specific tokenizers were used: PyThaiNLP (Phatthiyaphaibun et al., 2023) for Thai, MeCab for Japanese, and the provided zh tokenizer for Chinese. For reference-free quality estimation, we employ COMET-KIWI (Rei et al., 2022b).

5 Results

5.1 Transcript Translation Results

Table 3 reports macro-averaged results for transcript translation. We compare translation performance using aligned ASR and human-corrected transcripts under three context settings (Section 3.2): no additional context, aligned slide text, and aligned slide images. The aligned ASR condition evaluates translation from speech-derived inputs after edit-distance-based alignment, whereas the corrected transcript condition removes recognition errors from the source text. Corrected transcripts consistently yield higher ChrF, BLEU, and NIST scores than the former, with the exception of GPT-4.1 with slide images context for which NIST decreases from 6.3 to 5.8. This gap indicates that *ASR errors remain a major source of degradation* in lecture translation, even after segment-level alignment.

Among the evaluated models, GPT-5 mini achieves the strongest overall performance, obtaining the highest ChrF, BLEU, and NIST scores across all three context settings for both aligned ASR and corrected transcript inputs. Among open-weight models, Llama 4 Scout, Gemma 3 12B, and Qwen3.5 9B are the most competitive, with the strongest results depending on the input condition and metric. Llama 4 Scout leads on ChrF and BLEU for aligned ASR without additional context, Gemma 3 12B leads on BLEU with slide text, and Qwen3.5 9B leads on ChrF and BLEU with slide images.

Adding slide images produces mixed effects across models. GPT-5 mini remains relatively stable, while some models, particularly Llama 3.2 11B, degrade substantially when visual inputs are added, suggesting that current multimodal systems do not always use slide context effectively during translation.

Model	Aligned ASR					Corrected Transcript				
	ChrF	BLEU	COMET	KIWI	NIST	ChrF	BLEU	COMET	KIWI	NIST
no additional context										
Llama 4 Scout	56.8	37.9	83.9	76.1	6.0	63.1	44.5	87.3	81.4	6.6
Llama 3.2 11B	49.4	29.0	81.5	74.1	5.3	55.9	35.9	84.7	79.2	5.9
Gemma 3 12B	53.8	34.4	84.8	77.1	5.8	62.9	45.8	88.2	81.9	6.9
Qwen3.5 9B	55.7	36.5	85.1	77.4	6.0	60.7	41.5	88.1	82.2	6.4
Aya Vision 8B	51.8	34.1	80.9	72.9	5.5	57.0	39.4	84.2	78.1	6.0
InternVL 3.5 8B	49.9	30.6	79.9	73.0	5.1	55.4	36.1	83.6	78.5	5.6
GLM-4.6V Flash	52.6	33.4	81.7	73.8	5.6	58.2	39.3	85.4	79.4	6.1
GPT-4.1	55.8	36.5	86.5	78.2	6.0	66.9	49.7	89.0	82.5	7.1
GPT-5 mini	**62.4**	**45.7**	**87.4**	78.2	**6.8**	**71.3**	**57.7**	**90.4**	82.5	**7.9**
o4-mini	59.2	41.3	87.2	**78.2**	6.4	66.6	50.6	90.1	**82.6**	7.4
with slide text										
Llama 4 Scout	56.7	37.5	83.9	76.1	5.9	62.7	43.7	87.1	81.2	6.4
Llama 3.2 11B	49.1	28.7	81.1	73.6	5.2	54.8	34.9	84.3	78.6	5.8
Gemma 3 12B	56.7	38.3	84.9	77.0	6.2	63.0	46.1	88.0	81.7	6.9
Qwen3.5 9B	55.8	36.4	85.3	77.6	6.0	60.3	40.7	87.9	82.0	6.3
Aya Vision 8B	51.1	33.1	80.6	72.7	5.3	56.2	38.5	83.8	77.7	5.8
InternVL 3.5 8B	50.2	30.9	80.2	73.1	5.2	56.1	37.6	83.8	78.5	5.9
GLM-4.6V Flash	52.4	33.0	82.0	74.0	5.5	58.0	38.9	85.4	79.1	6.0
GPT-4.1	52.7	32.3	86.0	77.9	5.5	66.0	48.7	89.0	82.4	7.0
GPT-5 mini	**60.4**	**43.2**	**87.2**	**78.3**	**6.6**	**69.8**	**55.5**	**90.2**	82.4	**7.7**
o4-mini	56.4	38.6	84.8	76.6	5.9	65.7	49.5	89.8	**82.5**	7.3
with slide image										
Llama 4 Scout	57.1	38.4	83.8	76.0	6.1	62.7	44.2	87.1	81.0	6.6
Llama 3.2 11B	39.9	19.5	74.8	67.6	3.6	46.6	26.3	79.8	73.8	4.5
Gemma 3 12B	55.3	36.3	84.9	77.0	6.0	63.5	46.8	88.1	81.7	6.9
Qwen3.5 9B	58.0	40.0	85.1	77.2	6.3	63.1	45.2	88.0	81.9	6.7
Aya Vision 8B	50.9	33.4	80.5	72.6	5.4	56.1	39.0	83.7	77.7	5.9
InternVL 3.5 8B	50.3	31.4	80.3	73.4	5.3	55.6	37.0	83.8	78.5	5.8
GLM-4.6V Flash	51.3	30.8	79.0	72.9	5.2	55.7	35.2	82.8	78.1	5.5
GPT-4.1	57.6	39.0	86.8	78.3	6.3	58.4	39.8	83.5	76.5	5.8
GPT-5 mini	**62.2**	**45.3**	**87.5**	**78.3**	**6.8**	**70.1**	**55.8**	**90.2**	**82.5**	**7.8**
o4-mini	58.6	40.4	87.1	78.3	6.3	65.0	48.5	89.5	82.3	7.1

Table 3: Overall model performance on LECTRANS Transcript Translation task, reported as macro-averaged metrics. COMET scores are scaled by 100. **Bold** indicates highest, while underlined indicates second highest.

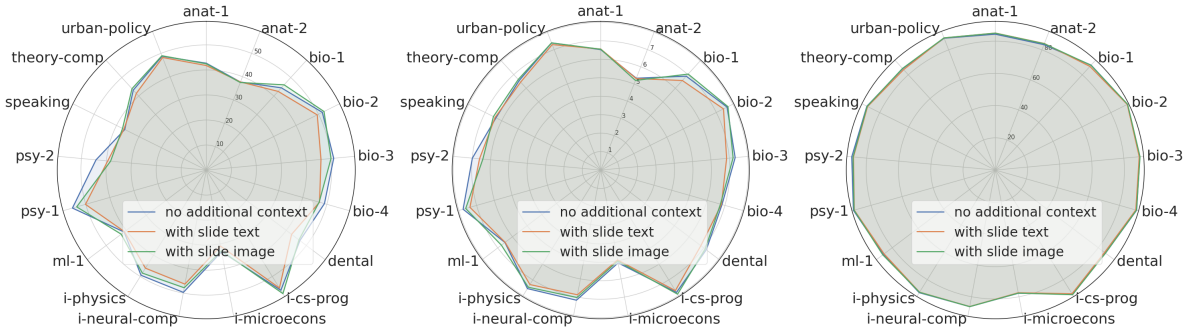
Model	*no-context*					*with slide image*				
	ChrF	BLEU	COMET	KIWI	NIST	ChrF	BLEU	COMET	KIWI	NIST
Llama 3.2 11B	58.4	22.5	72.1	73.6	3.0	35.4	9.5	69.8	66.8	1.3
Gemma 3 12B	70.5	49.7	81.9	76.8	5.7	71.1	42.2	86.1	76.7	4.9
GPT-4.1	74.6	63.7	92.7	79.7	7.1	75.2	64.8	92.8	79.6	7.1
GPT-5 mini	**75.7**	**64.7**	**92.9**	**80.0**	**7.1**	**76.3**	**65.8**	**93.0**	80.0	**7.2**
o4-mini	74.8	63.8	92.7	79.8	7.1	75.1	64.5	92.8	**80.0**	7.1

Table 4: Overall model performance on LECTRANS Slide Translation task, reported as averaged metrics. COMET scores are scaled by 100. **Bold** indicates highest, while underlined indicates second highest.

5.2 Slide Translation Results

Table 4 reports results on the slide translation task. We compare text-only translation of extracted slide text with image-grounded translation, where models additionally receive the original slide image. GPT-5 mini achieves the strongest overall performance across both settings, followed closely by GPT-4.1 and o4-mini. For these proprietary models, adding slide images yields small but generally consistent improvements across reference-based metrics.

Open-weight models show less stable behavior. Some models benefit from image grounding on selected metrics, such as Gemma 3 12B with COMET score increases 4.2 points, while others degrade substantially when full-slide visual context is added, particularly Llama 3.2 11B with BLEU falling from 22.5 in the text-only setting to 9.5 when slide image context is added. These results suggest that slide images can provide useful grounding, but may also distract models from the local text-region translation objective, especially in slides containing dense academic content.



(a) BLEU (b) NIST (c) COMET

Figure 4: Max scores per subject across models grouped by configuration of transcript translation task.

Overall, the results indicate that strong text-only translation is difficult to improve through simple slide-context concatenation or full-slide image inputs. This motivates the more detailed analysis in the following section, where we examine variation across context settings, lecture subjects, and reasoning token usage.

6 Analysis and Discussion

We focus on several key findings here and defer additional analysis to Appendix D.

Effects of Slide Context. The experimental results suggest that current multimodal models do not reliably use slide context in translation settings with loose cross-modal alignment. In transcript translation, aligned slide text provides only limited benefits, while slide images produce highly model-dependent behavior, including degradation for some models such as Llama 3.2 11B.

One possible explanation is that models fail to distinguish the primary translation target from the supplementary visual context. In transcript translation, slides are intended to ground the spoken segment rather than serve as additional text to translate. Because lecture slides may summarize previous material, introduce later content, or omit orally explained details, multimodal models may overfit to visible slide content and shift away from the transcript itself.

A related pattern appears in slide translation. Proprietary models generally benefit from full-slide visual context, whereas open-weight models show less stable behavior. Although slide images can provide useful layout and neighboring text cues, they may also distract models from local text-region translation. These results suggest that a more structured use of slide layout, bounding boxes, and element types may be more effective than simple text or image context prompting.

Performance by Lecture Subject. Figure 4 shows clear variation across lecture subjects. The `i-microecons` subject generally presents difficulties for transcript translation and has the highest aligned ASR word error rate, at 36.7% (Table 10). By contrast, `psy-1` achieves comparatively strong translation results and has an aligned word error rate of 9.1%. For example, GPT-5 mini without additional context obtains BLEU scores of 55.5 on `psy-1`, 41.1 on `ml-1`, and 33.2 on `i-microecons`. The results indicate that source transcript quality can strongly affect downstream translation performance, even after edit-distance-based alignment. Furthermore, based on our ASR error analysis in Appendix G, these results are consistent with source transcript quality contributing to subject-level variation.

Subject-level variation is not solely explained by ASR quality. Lecture domains differ in terminology density, mathematical notation, slide design, and the extent to which spoken explanations depend on visual content. Technical lectures with dense slide content or specialized notation can be challenging for both transcript and slide translation. This suggests that academic lecture translation should be evaluated across diverse lecture subjects rather than relying only on aggregate scores. Additional analysis by target language is provided in Appendix D, and complete subject-level results are reported in Appendix N.

Thinking token size and translation quality. Recent work suggests that reasoning in MMT is most effective when applied selectively, as in DeDIT, which invokes visual reasoning only for ambiguous inputs (Liu et al., 2025). This is particularly relevant for lecture translation, where slide

Metric	*no additional context*	*with slide text*	*with slide image*
ChrF	-.6188	.0315	-.6247
BLEU	-.5898	-.0060	-.6551
NIST	-.5345	.0460	-.5140
COMET	-.8051	.0925	-.5790
KIWI	-.7038	.1407	-.4812

Table 5: Pearson correlations between reasoning-token counts and translation-quality metrics for GPT-5 mini across the three transcript translation context settings.

context may either distract or ground the transcript. We therefore examine whether longer reasoning traces correspond to better translation quality.

Using GPT-5 mini, Table 5 reports Pearson correlations between reasoning-token counts and evaluation scores across the three transcript translation settings, while Figure 10 in Appendix K visualizes the relationship between reasoning-token usage and BLEU scores across subjects and configurations. In both the no-context setting and slide-image settings, reasoning-token counts show negative correlations with translation quality across metrics. By contrast, correlations in the slide-text setting are near-zero for ChrF, BLEU, and NIST metrics.

These results do not suggest that longer reasoning improves translation quality in this setting. Instead, longer reasoning traces may reflect uncertainty, difficult source segments, or unstable use of visual context. The difference between slide-text and slide-image settings also suggests that reasoning behavior depends on the form of multimodal input. Additional reasoning tokens should therefore not be assumed to improve translation quality in lecture translation.

7 Conclusion

We introduced LECTRANS, a benchmark for multimodal academic lecture translation constructed from more than 383 hours of English lecture videos. The dataset includes transcripts, slide images, extracted slide text, and an eight-language parallel evaluation subset. We formulate lecture translation as two tasks: transcript translation for spoken lecture content and slide translation for visually presented text.

Our experiments show that transcript quality strongly affects downstream performance. Human-corrected transcripts consistently improve ChrF and BLEU relative to aligned ASR inputs, indicating that ASR errors remain a major bottleneck for lecture translation. In both tasks, GPT-5 mini achieved the strongest overall performance on reference-based metrics, while several open-weight multimodal models remain competitive in selected settings.

The results further show that slide context is not uniformly beneficial. Although slide text and slide images can provide useful grounding, loosely aligned visual content may also distract models from the primary translation target. These findings suggest that effective lecture translation requires more than merely adding slide text or full-slide images as auxiliary context. Future work should explore more structured cross-modal integration, including slide layout, bounding boxes, temporal alignment, and terminology-aware evaluation.

Limitations

LECTRANS evaluates segment-level translation rather than complete end-to-end translation from raw lecture audio or video. Although the benchmark includes ASR outputs, slide images, OCR text, and aligned multimodal context, our experiments do not evaluate a full pipeline that starts from raw audiovisual input and produces final translated lecture content.

The current multimodal settings use simple slide text or slide image as context. They do not fully exploit slide structure, bounding-box categories, mathematical notation, diagrams, or temporal relations across multiple slides available in LECTRANS. More structured representations may be needed to determine when slide content supports transcript translation and when it acts as noise in cross-modal representations.

The human-verified translation subset is smaller than the full automatically processed corpus. The full collection contains more than 383 hours of lecture video, while the human-verified translation references are available only for selected segments in the 162-minute evaluation subset. This design allows careful verification across seven target languages, but it limits the scale of reference-based evaluation.

Finally, the evaluation relies mainly on automatic MT metrics. These metrics are useful for comparing systems, but they may not fully capture terminology consistency, pedagogical faithfulness, or whether a translation correctly uses evidence from the aligned slide context. Future evaluations should include more targeted measures for domain

terminology, cross-modal grounding, and learner-facing adequacy.

Ethical Considerations

LECTRANS is built from publicly available educational lecture resources. We do not redistribute the original videos or audio recordings, the release contains only derived materials, including metadata, aligned transcripts, extracted slide information, annotations, translations, and evaluation scripts. Users are responsible for complying with the licenses and terms of use of the original lecture providers, especially when accessing or processing source videos. We followed the relevant institutional guidelines for data collection, annotation, and benchmark construction.

Human correction and verification were performed by legally hired annotators and professional translators depending on the task. For Transcription Pipeline, we hired professional annotators: one native English speaker for ASR correction and seven native speakers of the respective languages for transcript translation. For Slide Pipeline, we hired graduate-student annotators: 18 students with English proficiency of CEFR B2 or higher for OCR correction and seven students who are native speakers of the languages studied for slide text translation. Annotators were compensated according to the relevant institutional and contractual arrangements, at hourly rates of approximately USD 35 for professional annotators and USD 9 for student annotators, paid in local currency. A total of more than 700 hours of annotation work has been conducted for this study. We have released our dataset and code, with the dataset format as described in Appendix H.

The benchmark may still contain errors from ASR, OCR, machine translation, or human correction, which can affect technical terminology and misrepresent academic content if used without review. We therefore frame LECTRANS as an evaluation benchmark rather than a deployable lecture translation service.

For transparency, we report the computational cost here. We use around 500 GPU-hours for Llama 4 Scout on NVIDIA RTX Pro 6000 GPU. We use about another 500 GPU-hours for other open-weight models including the inference and evaluations on NVIDIA Ampere and Ada GPUs. We also use approximately USD 450 in OpenAI credits for proprietary model experiments, and data pre-annotations.

AI assistants were used for language editing with a privacy-preserving tool, following the Publication Ethics Committee policy.

Acknowledgments

This work is supported by JSPS KAKENHI under Grant Number JP21H05054 and JST SPRING under Grant Number JPMJSP2140.

References

Ana Bell. 2022. 6.100L Introduction to CS and Programming using Python, Fall 2022. MIT OpenCourseWare. Accessed: Aug 27, 2025.

Markus Buehler and Jeffrey Grossman. 2012. 3.021J Introduction to Modeling and Simulation, Spring 2012. MIT OpenCourseWare. Accessed: Nov 13, 2025.

Zhe Chen, Heyang Liu, Wenyi Yu, Guangzhi Sun, Hongcheng Liu, Ji Wu, Chao Zhang, Yu Wang, and Yanfeng Wang. 2024. M^{3}AV: A multimodal, multigenre, and multipurpose audio-visual academic lecture dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9041–9060, Bangkok, Thailand. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, page 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Kosuke Doi, Mana Makinae, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2026. SimulCOMET: A quality metric for simultaneous interpretation in distant language pair considering word order

difference. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 42515–42533, San Diego, California, United States. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Michale Fee and Daniel Zysman. 2018. 9.40 Introduction to Neural Computation, Spring 2018. MIT OpenCourseWare. Accessed: Aug 27, 2025.

Yi Feng, Chuanyi Li, Jiatong He, Zhenyu Hou, and Vincent Ng. 2025. Multimodal neural machine translation: A survey of the state of the art. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22130–22147, Suzhou, China. Association for Computational Linguistics.

Christian Fügen, Muntsin Kolss, Dietmar Bernreuther, Matthias Paulik, Sebastian Stüker, Stephan Vogel, and Alexander H. Waibel. 2006. Open domain speech recognition & translation:lectures and speeches. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I.

John D. E. Gabrieli. 2011. 9.00SC Introduction to Psychology, Fall 2011. MIT OpenCourseWare. Accessed: Nov 13, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Eric Grimson, John Guttag, and Ana Bell. 2016. 6.0002 Introduction to Computational Thinking and Data Science, Fall 2016. MIT OpenCourseWare. Accessed: Nov 13, 2025.

Conghui He, Wei Li, Zhenjiang Jin, Chao Xu, Bin Wang, and Dahua Lin. 2024. Opendatalab: Empowering general artificial intelligence with open datasets. *Preprint*, arXiv:2407.13773.

David Hsu. 2022. 11.165J Urban Energy Systems and Policy, Fall 2022. MIT OpenCourseWare. Accessed: Nov 13, 2025.

David Kaiser. 2020. STS.042J / 8.225J Einstein, Oppenheimer, Feynman: Physics in the 20th Century, Fall 2020. MIT OpenCourseWare. Accessed: Aug 27, 2025.

Sai Koneru, Fabian Retkowski, Christian Huber, Lukas Hilgert, Seymanur Akti, Enes Yavuz Ugan, Alexander Waibel, and Jan Niehues. 2026. BOOM: Beyond only one modality KIT's multimodal multilingual lecture companion. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 175–187, Rabat, Marocco. Association for Computational Linguistics.

Dong Won Lee, Chaitanya Ahuja, Paul Pu Liang, Sanika Natu, and Louis-Philippe Morency. 2023. Lecture presentations multimodal dataset: Towards understanding multimodality in educational videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20087–20098.

Danyang Liu, Fanjie Kong, Xiaohang Sun, Dhruva Patil, Avijit Vajpayee, Zhu Liu, Vimal Bhat, and Najmeh Sadoughi. 2025. Detect, disambiguate, and translate: On-demand visual reasoning for multimodal machine translation with large vision-language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1559–1570, Albuquerque, New Mexico. Association for Computational Linguistics.

Mana Makinae, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024. Simul-MuST-C: Simultaneous multilingual speech translation corpus using large language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22185–22205, Miami, Florida, USA. Association for Computational Linguistics.

Massachusetts Institute of Technology. 2001. MIT OpenCourseWare. `https://ocw.mit.edu`.

Meta. 2025. meta-llama/llama-4-scout-17b-16e-instruct. `https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct`. HuggingFace Model card. Accessed May-25-2026.

Markus Müller, Thai Son Nguyen, Jan Niehues, Eunah Cho, Bastian Krüger, Thanh-Le Ha, Kevin Kilgour, Matthias Sperber, Mohammed Mediani, Sebastian Stüker, and Alex Waibel. 2016. Lecture translator - speech translation framework for simultaneous lecture translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 82–86, San Diego, California. Association for Computational Linguistics.

Wannaphong Phatthiyaphaibun, Korakot Chaovavanich, Charin Polpanumas, Arthit Suriyawongkul, Lalita Lowphansirikul, Pattarawat Chormai, Peerat Limkonchotiwat, Thanathip Suntorntip, and Can Udomcharoenchaikit. 2023. PyThaiNLP: Thai natural language processing in python. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 25–36, Singapore. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents. `https://huggingface.co/Qwen/Qwen3.5-9B`. Cited from Hugging Face model card information. Accessed: May-25-2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Huangjun Shen, Liangying Shao, Wenbo Li, Zhibin Lan, Zhanyu Liu, and Jinsong Su. 2024. A survey on multi-modal machine translation: Tasks, methods and challenges. *Preprint*, arXiv:2405.12669.

Michael Sipser. 2020. 18.404J Theory of Computation, Fall 2020. MIT OpenCourseWare. Accessed: Aug 27, 2025.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.

Matthias Sperber and Matthias Paulik. 2020. Speech translation and the end-to-end promise: Taking stock of where we are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7409–7421, Online. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 74 others. 2026. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Ayu Teramen, Takumi Ohtsuka, Risa Kondo, Tomoyuki Kajiwara, and Takashi Ninomiya. 2024. English-to-Japanese multimodal machine translation based on image-text matching of lecture videos. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 86–91, Bangkok, Thailand. Association for Computational Linguistics.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Robert M. Townsend. 2020. 14.04 Intermediate Microeconomic Theory, Fall 2020. MIT OpenCourseWare. Accessed: Nov 13, 2025.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *Preprint*, arXiv:2502.14786.

Hoang-Son Vo, Karina Kolmogortseva, Ngumimi Karen Iyortsuun, Hong-Duyen Vo, and Soo-Hyung Kim. 2025. Globalizeed: A multimodal translation system that preserves speaker identity in academic lectures. *Preprint*, arXiv:2510.11596.

Jiaqi Wang, Ricky Y.-K. Kwok, and Edith C. H. Ngai. 2025a. Towards key point identification (kpi) for lecture videos: Approaches and performance evaluation. *ACM Trans. Multimedia Comput. Commun. Appl.*, 21(7).

Model Name	Model ID
Llama 4 Scout	meta-llama/Llama-4-Scout-17B-16E-Instruct
Llama 3.2 11B	meta-llama/Llama-3.2-11B-Vision-Instruct
Gemma 3 12B	google/gemma-3-12b-it
Qwen3.5 9B	Qwen/Qwen3.5-9B
Aya Vision 8B	CohereLabs/aya-vision-8b
InternVL 3.5 8B	OpenGVLab/InternVL3_5-8B-HF
GLM-4.6V Flash	zai-org/GLM-4.6V-Flash
GPT-4.1	gpt-4.1-2025-04-14
GPT-5 mini	gpt-5-mini-2025-08-07
o4 mini	o4-mini-2025-04-16

Table 6: Actual Model ID used in this study

Minghan Wang, Yuxia Wang, Thuy-Trang Vu, Ehsan Shareghi, and Reza Haf. 2024. Exploring the potential of multimodal LLM with knowledge-intensive multimodal ASR. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13274–13288, Miami, Florida, USA. Association for Computational Linguistics.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025b. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Ryoya Yuasa, Akihiro Tamura, Tomoyuki Kajiwara, Takashi Ninomiya, and Tsuneo Kato. 2023. Multimodal neural machine translation using synthetic images transformed by latent diffusion model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 76–82, Toronto, Canada. Association for Computational Linguistics.

A Languages Studied

We translate the English sources into seven target languages selected to cover diverse linguistic backgrounds. To illustrate language relatedness among the eight languages involved, we present their family trees as shown in Figure 5.

B Model List

We employ OpenAI's batch API for the three GPT models used throughout our experiments. The remaining models are all open-sourced models, and we utilize HuggingFace library. The actual model identifiers used in this study are listed in Table 6. We use only the default parameter settings.

C Prompt for Translation Task

We show the prompt used throughout our experiments as follows.

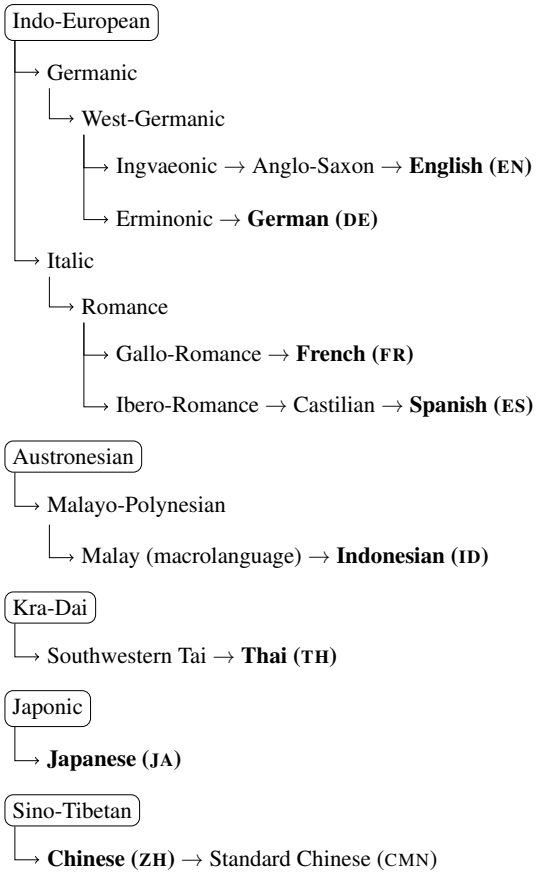


Figure 5: Family trees of eight languages included in LECTRANS. **Bold** represents the actual language used.

Transcript Transl. No Additional Context

You are a translation system for lecture video transcripts. Translate the spoken transcript into the specified target language.

Translate the following from English to `{target_language}`. Please return only the translation.

English:
`"{transcript_en}"`

`{target_language}`:

Transcript Transl. With Slide Text

You are a translation system for lecture video transcripts. Translate the spoken transcript into the specified target language.

In addition to the transcript, you are given auxiliary textual context extracted from the corresponding presentation slides.

This context is provided to supplement semantic understanding and disambiguate technical terms, but it is not part of the content to be translated.

The context is as follows:
`{slides_text}`

Translate the following from English to `{target_language}`. Please return only the translation.

English:
`"{transcript_en}"`

`{target_language}`:

Transcript Transl. With Slide Image

You are a translation system for lecture video transcripts. Translate the spoken transcript into the specified target language.

In addition to the transcript, you are given auxiliary image context from the corresponding presentation slides.

This context is provided to supplement semantic understanding and disambiguate technical terms, but it is not part of the content to be translated.

`{slides_image}`

Translate the following from English to `{target_language}`. Please return only the translation.

English:
`"{transcript_en}"`

`{target_language}`:

Slide Translation (both config)

You are a translation system for academic lecture slides. Translate only the text contained in each bounding box into the specified target language. You will be given the slide image from which the bounding boxes were extracted, when available.

The input will always be a structured JSON object. Each key is a bounding-box ID (a string in double-quotes). Each value is the text to translate, enclosed in triple double-quotes (""").

Your task:
Translate only the text inside the triple-quoted strings into the specified target language.

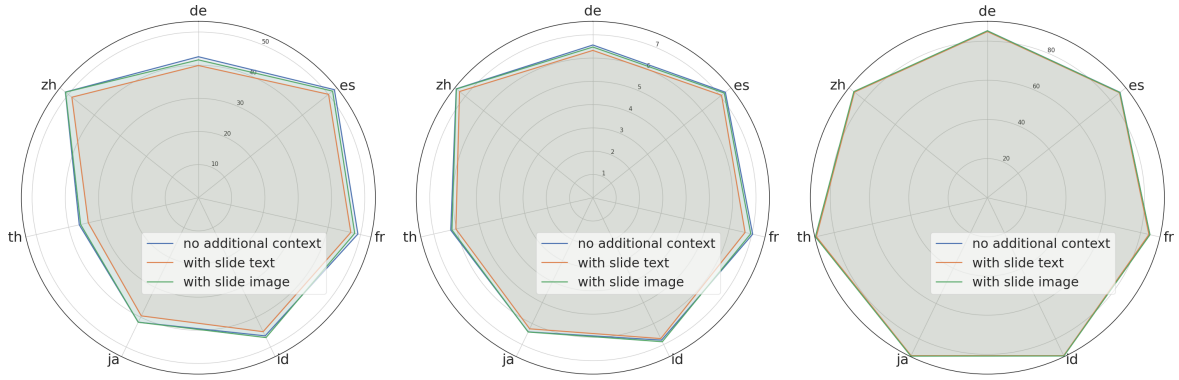
Rules:
1. Keep the JSON structure exactly the same.
2. Keep all keys (bounding-box IDs) exactly as given.
3. Replace only the text inside each triple-quoted value with its translation.
4. Do not output any commentary, explanations, or extra text. Return only the translated JSON object.

{
"target_language": "`{lang_tgt}`",
"bounding_boxes": { `{input_src}` }
}

D Analysis and Discussion (Continued)

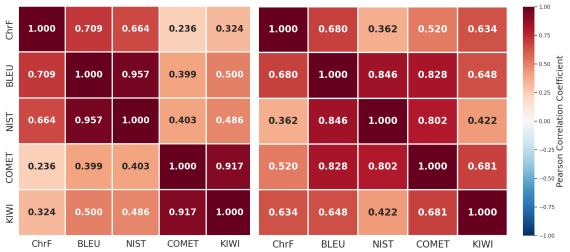
Performance by Target Language. Figure 6 shows that transcript translation performance differs substantially across target languages. The highest scores are generally obtained for Spanish, French, German, and Indonesian, while Japanese and Thai tend to receive lower lexical scores. Chinese shows a more metric-dependent pattern: it obtains competitive BLEU and NIST scores in several settings, but lower ChrF than languages written in Latin script. This suggests that language-level comparisons should be interpreted with care, especially when metrics differ in their sensitivity to script, tokenization, and character-level overlap.

These differences likely reflect several factors rather than a single cause. Target languages differ in word order, morphology, segmentation behavior, and script. In addition, multilingual and multimodal models may have uneven coverage across languages in their pre-training and instruction-tuning data. The results therefore show that lecture translation remains uneven across target languages, even when the same source content and model configuration are used. For the complete result see Appendix L.



(a) BLEU (b) NIST (c) COMET

Figure 6: Max scores per language across models grouped by configuration of transcript translation task.



(a) Transcripts (b) Slides

Figure 7: Correlation among the five metrics used in translation task.

COMET-based metrics show different trends. Figure 7 shows that the evaluation metrics capture different aspects of translation quality. ChrF, BLEU, and NIST correlate strongly with one another, reflecting their shared reliance on surface-form and n-gram overlap. COMET and COMET-KIWI also correlate strongly with each other with $r = 0.917$, but their correlation with lexical metrics is weaker. The metrics correlations for Transcript Task and Slide Task is as shown by Figure 7a and Figure 7b, respectively. This divergence affects model ranking. For transcript translation, GPT-5 mini performs best on all metrics except COMET-KIWI, while Gemma 3 12B and Qwen 3.5 9B obtain notably high COMET and COMET-KIWI scores. This suggests that neural metrics may reward semantic adequacy differently from lexical metrics. Conversely, lexical metrics may be more sensitive to exact wording and less frequent n-gram matches, which are relevant in technical lecture content but do not directly measure terminology accuracy. A complete evaluation of academic lecture translation therefore requires both standard MT metrics and more targeted analyses of terminology, grounding, and pedagogical faithfulness.

Subject	Video Length	Retrieved Slide #	# after De-duplicated	Avg. Secs. per Slide
anat-1	0:44:37	909	50	53.5
anat-2	0:27:12	537	16	102.0
bio-1	0:50:53	270	47	65.0
bio-2	0:08:44	75	9	58.2
bio-3	0:43:52	240	57	46.2
bio-4	0:30:34	603	38	48.3
dental	0:22:37	222	19	71.4
i-microecons	0:59:06	597	77	46.1
i-cs-prog	0:47:47	1725	75	38.2
i-neural-comp	1:18:55	1152	117	40.5
i-physics	1:20:21	756	50	96.4
ml-1	1:09:35	2103	114	36.6
psy-1	0:51:37	324	20	154.9
psy-2	0:36:25	372	18	121.4
speaking	0:56:50	1467	84	40.6
theory-comp	1:17:35	3255	101	46.1
urban-policy	0:33:45	1191	60	33.8

Table 7: Slide extraction statistics.

E Slide Extraction Statistics

Please refer to Table 7.

F Performance using Raw ASR as Translation Input

Please refer to Table 8 and Table 9.

G Alignment of ASR to Human Annotation

Table 10 summarizes the discrepancy between the raw ASR outputs and the ASR outputs that is aligned to human-corrected transcripts. Table 11 reports sequence lengths, minimum edit distance, and the number of possible string alignment operations attaining the minimum edit distance. Given the astronomical amount of path count, we considered two edit-distance libraries to help us determine the path, namely, python-Levenshtein and Edlib. The

Model	Raw ASR				
	ChrF	BLEU	COMET	KIWI	NIST
no additional context					
Llama 4 Scout	48.6	30.5	72.8	61.8	4.9
Llama 3.2 11B	43.3	24.4	71.7	61.3	4.5
Gemma 3 12B	46.5	28.5	74.7	64.0	4.9
GPT-4.1	47.5	29.5	76.4	**65.4**	5.0
GPT-5 mini	**53.2**	**37.8**	**76.8**	65.3	**5.7**
o4-mini	50.5	33.8	76.4	65.0	5.4
with slide text					
Llama 4 Scout	48.2	29.8	73.3	63.0	4.8
Llama 3.2 11B	42.3	23.2	71.7	61.6	4.3
Gemma 3 12B	48.6	31.6	74.5	63.7	5.2
GPT-4.1	45.8	27.0	**77.5**	**67.1**	4.8
GPT-5 mini	**51.4**	**35.2**	77.0	65.9	**5.5**
o4-mini	48.1	31.7	75.5	64.7	5.0
with slide image					
Llama 4 Scout	49.0	31.1	73.1	62.2	5.0
Llama 3.2 11B	33.9	14.5	66.2	57.7	2.8
Gemma 3 12B	47.7	30.0	75.0	64.2	5.0
GPT-4.1	48.9	31.7	76.6	65.6	5.3
GPT-5 mini	**53.1**	**37.2**	**77.6**	**66.1**	**5.7**
o4-mini	50.3	33.4	76.7	65.5	5.4

Table 8: Overall model performance on LECTRANS Transcript Translation task using *raw ASR output* as the input, reported as macro-averaged metrics. COMET are scaled by 100. **Bold** indicates highest, while underlined indicates second highest.

Model	Raw ASR		Aligned		Corrected	
	ChrF	BLEU	ChrF	BLEU	ChrF	BLEU
no additional context						
Llama 4 Scout	48.6	30.5	56.8	37.9	63.1	44.5
Llama 3.2 11B	43.3	24.4	49.4	29.0	55.9	35.9
Gemma 3 12B	46.5	28.5	53.8	34.4	62.9	45.8
GPT-4.1	47.5	29.5	55.8	36.5	66.9	49.7
GPT-5 mini	**53.2**	**37.8**	**62.4**	**45.7**	**71.3**	**57.7**
o4-mini	50.5	33.8	59.2	41.3	66.6	50.6
with slide text						
Llama 4 Scout	48.2	29.8	56.7	37.5	62.7	43.7
Llama 3.2 11B	42.3	23.2	49.1	28.7	54.8	34.9
Gemma 3 12B	48.6	31.6	56.7	38.3	63.0	46.1
GPT-4.1	45.8	27.0	52.7	32.3	66.0	48.7
GPT-5 mini	**51.4**	**35.2**	**60.4**	**43.2**	**69.8**	**55.5**
o4-mini	48.1	31.7	56.4	38.6	65.7	49.5
with slide image						
Llama 4 Scout	49.0	31.1	57.1	38.4	62.7	44.2
Llama 3.2 11B	33.9	14.5	39.9	19.5	46.6	26.3
Gemma 3 12B	47.7	30.0	55.3	36.3	63.5	46.8
GPT-4.1	48.9	31.7	57.6	39.0	58.4	39.8
GPT-5 mini	**53.1**	**37.2**	**62.2**	**45.3**	**70.1**	**55.8**
o4-mini	50.3	33.4	58.6	40.4	65.0	48.5

Table 9: Comparison of various transcriptions as inputs. **Bold** indicates highest, while underlined indicates second highest.

importance of applying the string alignment is reflected from the error rate as shown in Table 12. Terminology-wise ASR correctness is as reflected in Table 13

H Actual Dataset Format

This section details the format of our dataset. We do not provide audio or video material itself but rather example code on how to source the media.

Subject	WER↓ (%)		
	Raw ASR	Aligned	Document
anat-1	39.0	16.5	16.5
anat-2	34.0	24.4	24.4
bio-1	26.7	5.5	5.5
bio-2	21.9	5.5	5.5
bio-3	36.0	21.7	21.5
bio-4	28.6	15.7	15.4
dental	20.9	14.9	14.9
i-microecons	43.7	36.7	36.4
i-cs-prog	28.6	6.9	6.9
i-neural-comp	24.9	15.1	15.1
i-physics	28.2	10.1	10.1
ml-1	26.9	17.4	17.4
psy-1	35.2	9.1	9.1
psy-2	26.3	12.0	12.0
speaking	24.1	20.8	20.8
theory-comp	36.6	14.6	14.6
urban-policy	25.2	13.0	13.0

Table 10: Percentage of WER (Word Error Rate) of the ASR outputs against human-corrected transcripts for each lecture subject in LECTRANS. Raw ASR indicates the original timestamp-level Whisper outputs. Aligned indicates the raw ASR text after edit-distance-based string alignment to the human-corrected transcript segmentation, without any correction to the recognition errors. Document indicates the score after concatenating the raw ASR transcript in chronological order. For full error rate including CER (Character Error Rate) and library comparisons, the complete results are reported in the Appendix G on Table 12.

Subject	ASR Length	Annot. Length	Min. Dist.	# of Path w/ Min. Dist.
anat-1	9,013	8,401	1,125	7.96×10^{58}
anat-2	5,869	4,670	1,258	1.85×10^{55}
bio-1	9,961	9,702	359	1.24×10^{17}
bio-2	9,165	8,989	346	4.15×10^{21}
bio-3	6,622	6,281	950	1.66×10^{66}
bio-4	6,368	6,061	708	3.34×10^{32}
dental	8,332	7,509	875	6.55×10^{38}
i-microecons	7,616	5,681	2,085	1.75×10^{103}
i-cs-prog	8,072	7,655	441	3.02×10^{20}
i-neural-comp	7,914	7,290	873	8.12×10^{44}
i-physics	9,909	9,440	731	1.45×10^{30}
ml-1	7,062	7,214	513	6.39×10^{42}
psy-1	8,805	8,812	392	1.72×10^{20}
psy-2	8,442	7,994	585	7.75×10^{32}
speaking	8,356	7,766	829	9.16×10^{33}
theory-comp	6,347	5,946	578	3.12×10^{41}
urban-policy	8,473	7,931	681	3.76×10^{31}

Table 11: String alignment statistics using edit distance algorithm between ASR output and human-corrected transcripts (denoted as Annot.).

- Lecture video: raw lecture video data, we provided an example code to source.
- Lecture audio: audio extracted from the lecture videos.

Subject	Raw ASR				Aligned python-Levenshtein				Aligned Edlib			
	Segment		Document		Segment		Document		Segment		Document	
	CER	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER	WER
anat-1	0.354	0.390	0.125	0.165	0.126	0.165	0.125	0.165	0.126	0.165	0.125	0.165
anat-2	0.313	0.340	0.214	0.244	0.216	0.244	0.215	0.244	0.216	0.244	0.214	0.244
bio-1	0.270	0.267	0.036	0.055	0.036	0.055	0.036	0.055	0.036	0.055	0.036	0.055
bio-2	0.212	0.219	0.038	0.055	0.038	0.055	0.038	0.055	0.032	0.049	0.031	0.055
bio-3	0.296	0.360	0.143	0.215	0.144	0.217	0.143	0.217	0.144	0.217	0.143	0.217
bio-4	0.260	0.286	0.111	0.154	0.112	0.157	0.111	0.154	0.112	0.157	0.111	0.154
dental	0.165	0.209	0.105	0.149	0.106	0.149	0.105	0.149	0.106	0.149	0.105	0.149
i-microecons	0.346	0.437	0.274	0.364	0.275	0.367	0.273	0.367	0.275	0.366	0.273	0.366
i-cs-prog	0.294	0.286	0.055	0.069	0.055	0.069	0.055	0.069	0.055	0.069	0.055	0.069
i-neural-comp	0.212	0.249	0.110	0.151	0.111	0.151	0.110	0.151	0.111	0.151	0.110	0.151
i-physics	0.270	0.282	0.074	0.101	0.074	0.101	0.074	0.101	0.074	0.100	0.074	0.100
ml-1	0.180	0.269	0.071	0.174	0.071	0.174	0.071	0.174	0.071	0.174	0.071	0.174
psy-1	0.313	0.352	0.045	0.091	0.045	0.091	0.045	0.091	0.045	0.091	0.045	0.091
psy-2	0.223	0.263	0.069	0.120	0.070	0.120	0.069	0.120	0.070	0.120	0.069	0.120
speaking	0.137	0.241	0.099	0.208	0.099	0.208	0.099	0.208	0.099	0.208	0.099	0.208
theory-comp	0.310	0.366	0.091	0.146	0.092	0.146	0.091	0.146	0.092	0.146	0.091	0.146
urban-policy	0.206	0.252	0.080	0.130	0.081	0.130	0.080	0.130	0.081	0.130	0.080	0.130

Table 12: CER (Character Error Rate) and WER (Word Error Rate) of the ASR outputs against corrected transcriptions for each lecture subject in LECTRANS. Raw ASR indicates the original ASR outputs, while Aligned indicates the ASR text that is aligned to the corrected transcription using edit distance algorithm library: python-Levenshtein and Edlib. Both segment-wise error rate, where texts are segmented based on the timestamp, and document-wise error rate, which concatenated the whole text in a chronological order, are assessed.

Subject	# Uniq.	# Occur.	# Corr.	Accuracy
anat-1	37	125	68	54.4%
anat-2	25	50	31	62.0%
bio-1	29	91	65	71.4%
bio-2	28	105	82	78.1%
bio-3	16	30	22	73.3%
bio-4	5	9	7	77.8%
dental	76	161	136	84.5%
i-microecons	19	39	21	53.8%
i-cs-prog	7	9	5	55.6%
i-neural-comp	7	21	12	57.1%
i-physics	9	17	6	35.3%
ml-1	8	12	5	41.7%
psy-1	12	53	38	71.7%
psy-2	15	29	21	72.4%
speaking	9	18	14	77.8%
theory-comp	13	30	22	73.3%
urban-policy	10	15	9	60.0%
Total	325	814	564	69.3%

Table 13: Terminology recognition accuracy of the ASR model on the Testset - Translation subset. The macro-averaged Accuracy is 64.7%.

- Segmented lecture video: lecture videos segmented into scenes using PySceneDetect.
- Lecture slides: slide images extracted simultaneously with the segmented lecture videos.
- Raw ASR transcript: automatically generated transcripts obtained from the lecture audio.
- Aligned ASR transcript: raw ASR transcripts re-segmented to match the corrected transcript structure using edit-distance alignment.
- Corrected transcript: manually corrected versions of the ASR transcripts.
- Bounding boxes information: JSON files containing bounding box information in slide images.

I Annotated Dataset Samples

Figure 8 shows a sample of ASR correction performed by human.

J Slide Annotation Tool Platform

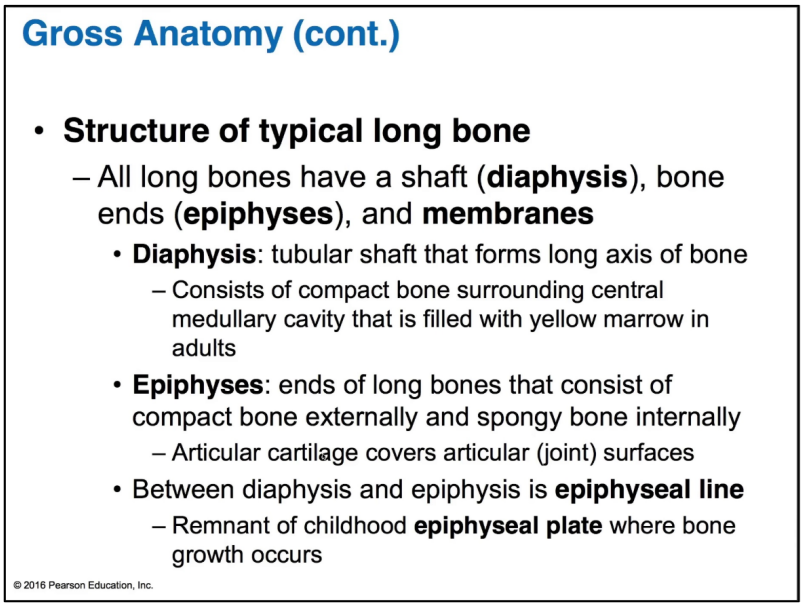
We use Label Studio Community Edition throughout the slide annotation process. Figure 9 shows a screenshot of this process.

K Reasoning Token and Performance Correlations

See Figures 10, 11, 12, 13, 14, and 15.

L Detailed Transcript Translation Performance By Target Language

The translation performances on all models grouped by target language directions on ChrF metrics can be found on Table 14, for BLEU on Table 15, COMET on Table 16, quality estimation of



~~Long bone~~ [OMITTED (OK)] All long bones, they have a shaft. The shaft is called a ~~diathesis~~ diaphysis [MISSPELL (CRIT)] [TERM]. Okay, ~~diathesis~~ diaphysis [MISSPELL (CRIT)] [TERM].

Figure 8: Correction of ASR output from `whisper-large-v3`, we outsourced the annotation to a third party professional annotator. We give relevant slides for the faithfulness of the annotation.

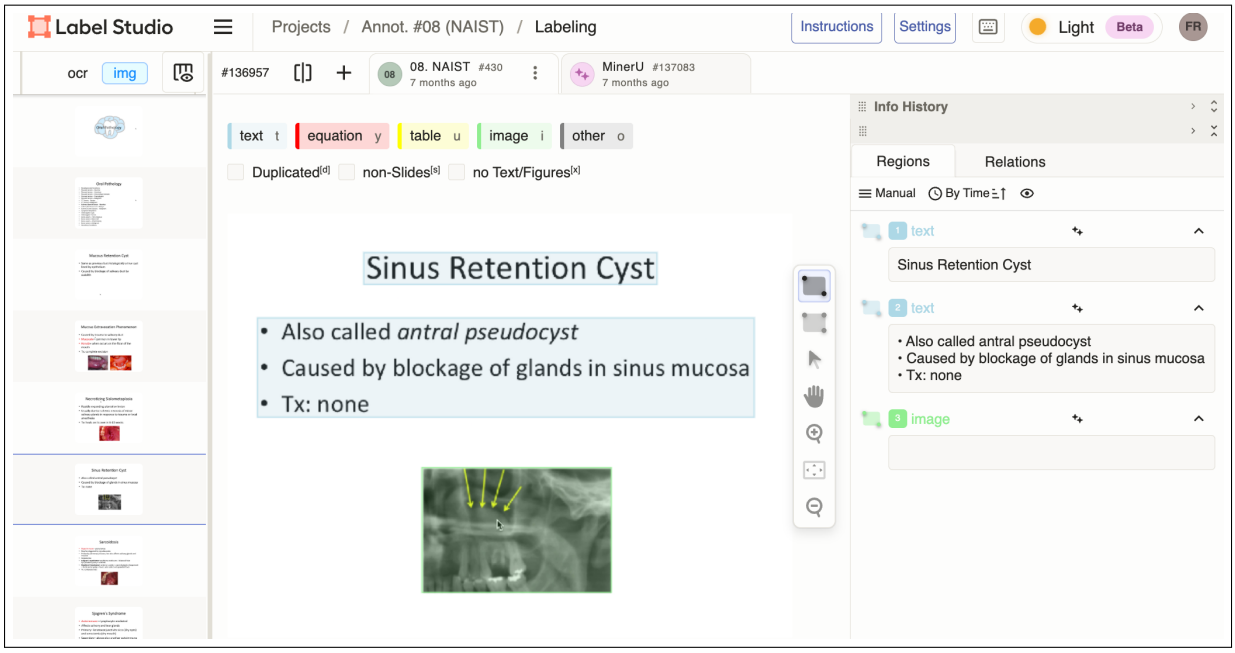


Figure 9: Sample screenshot showing annotation process of OCR correction on Label Studio Community Edition.

Transcript Transl.	ChrF (EN → ∗) JA	ZH	TH	ID	ES	FR	DE
no additional context							
Llama 4 Scout	35.5	33.9	53.3	67.5	**72.2**	69.7	65.2
Llama 3.2 11B	27.2	27.3	41.8	61.1	65.9	63.9	58.3
Gemma 3 12B	34.8	33.4	53.0	65.8	67.2	63.0	59.7
Qwen3.5 9B	36.5	33.6	55.2	67.8	69.2	66.3	61.6
Aya Vision 8B	34.8	33.8	27.4	66.7	69.7	68.2	62.0
InternVL 3.5 8B	24.6	37.9	36.2	61.5	66.4	64.9	57.8
GLM-4.6V Flash	32.7	38.2	42.0	60.2	69.3	66.3	59.5
GPT-4.1	40.8	36.5	54.0	65.3	67.3	65.1	61.6
GPT-5 mini	**47.8**	**46.3**	**62.0**	**70.7**	72.2	**70.3**	**67.6**
o4-mini	44.1	41.9	58.5	68.8	70.3	66.7	64.2
with slide text							
Llama 4 Scout	35.4	33.3	53.5	67.8	**71.7**	**70.2**	65.1
Llama 3.2 11B	26.2	26.6	42.1	60.6	66.4	64.2	57.7
Gemma 3 12B	37.2	35.2	54.8	68.6	70.9	67.4	62.5
Qwen3.5 9B	36.7	33.8	55.1	67.8	69.3	66.5	61.8
Aya Vision 8B	34.1	34.6	23.8	66.8	69.6	67.5	61.1
InternVL 3.5 8B	26.0	37.7	36.4	61.1	66.6	65.3	58.5
GLM-4.6V Flash	33.0	38.9	41.7	61.1	69.2	66.3	56.7
GPT-4.1	38.7	34.0	50.8	62.0	63.8	61.5	58.3
GPT-5 mini	**45.6**	**43.6**	**60.0**	**69.7**	70.6	67.7	**65.6**
o4-mini	40.6	37.8	54.7	65.8	68.6	64.6	63.0
with slide image							
Llama 4 Scout	36.0	34.6	53.5	68.0	**72.1**	**70.0**	65.3
Llama 3.2 11B	11.6	12.7	28.1	54.1	62.4	58.5	52.0
Gemma 3 12B	35.5	34.6	53.2	67.8	69.2	65.2	61.4
Qwen3.5 9B	37.5	37.8	56.5	69.6	71.7	68.9	64.4
Aya Vision 8B	34.1	34.5	22.6	66.6	69.3	67.9	61.5
InternVL 3.5 8B	27.7	37.7	34.2	62.8	67.1	64.8	57.6
GLM-4.6V Flash	31.0	37.4	41.7	59.2	67.5	62.9	59.3
GPT-4.1	42.3	38.1	55.2	67.2	69.6	67.3	63.6
GPT-5 mini	**47.8**	**46.3**	**61.7**	**71.0**	71.9	69.8	**67.0**
o4-mini	43.5	41.4	57.6	68.3	70.0	65.2	64.1

Table 14: Average ChrF score per language in Translation of Aligned ASR Transcripts.

COMET-KIWI on Table 17, and for assessing terminology co-occurrence using NIST on Table 18.

M Detailed Slide Translation Performance By Target Language

Please refer to Table 19

N Detailed Transcript Translation Performance By Subject

Translation performances on all models grouped by seventeen lecture subjects of LECTRANS against ChrF metrics can be found on Table 20, BLEU on Table 21, COMET on Table 22, quality estimation of COMET-KIWI on Table 23, and for assessing terminology co-occurrence using NIST on Table 24.

O Detailed Slide Translation Performance By Subject

Please refer to Table 25 and Table 26.

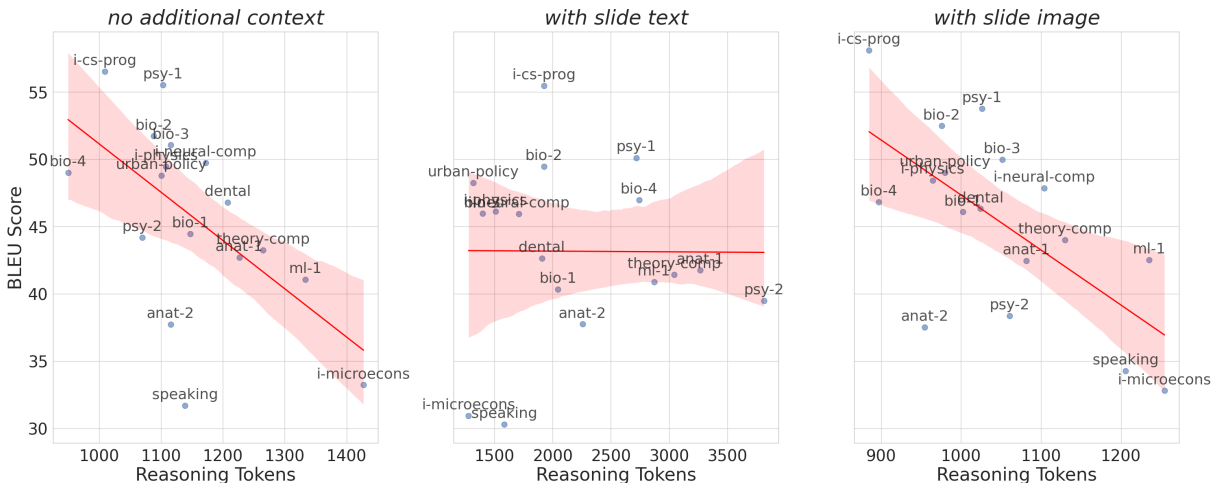


Figure 10: Relationship between Reasoning Tokens and BLEU Score by Subject and Config Separated using GPT-5 mini in transcript translation.

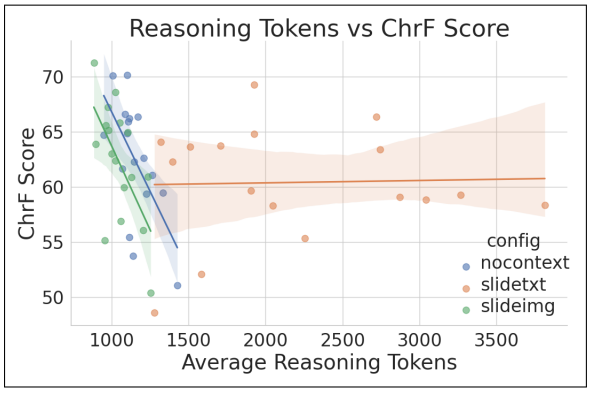


Figure 11: Reasoning Tokens vs average ChrF Score by subject on GPT-5 mini in transcript translation.

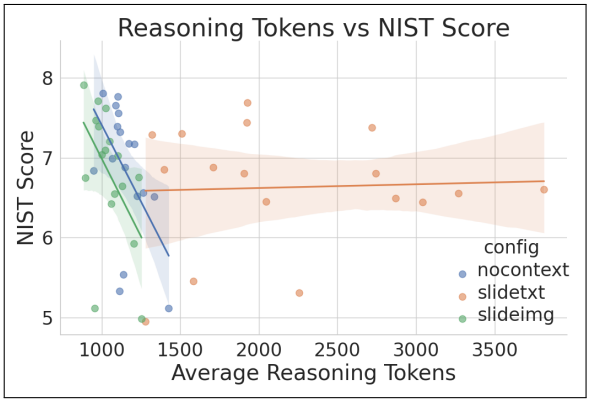


Figure 13: Reasoning Tokens vs average NIST Score by subject on GPT-5 mini in transcript translation.

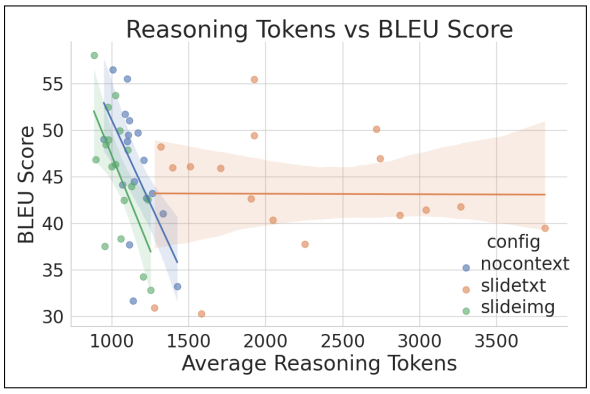


Figure 12: Reasoning Tokens vs average BLEU Score by subject on GPT-5 mini in transcript translation.

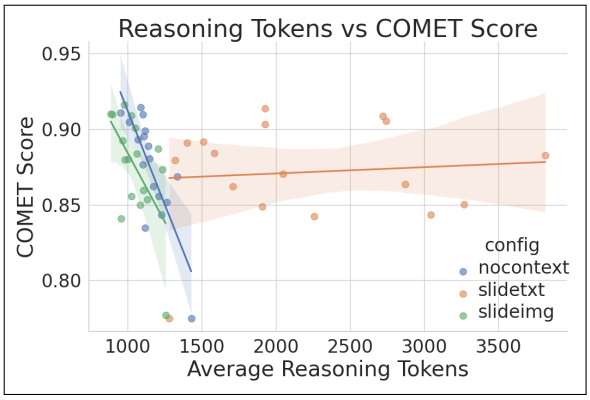


Figure 14: Reasoning Tokens vs average COMET Score by subject on GPT-5 mini in transcript translation.

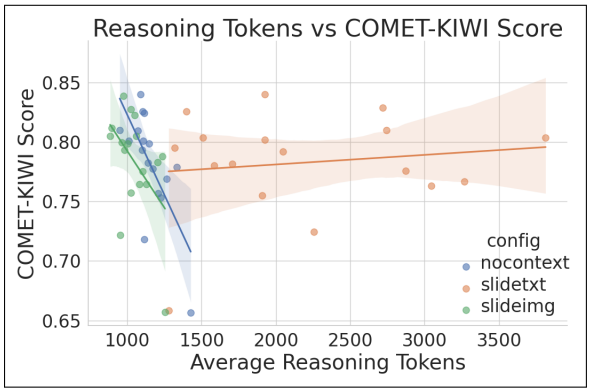


Figure 15: Reasoning Tokens vs average COMET-KIWI Score by subject on GPT-5 mini in transcript translation.

Transcript / Transl.	JA	ZH	TH	ID	ES	FR	DE
			BLEU (EN → ∗)				
no additional context							
Llama 4 Scout	27.8	38.1	24.8	40.0	50.5	46.2	37.7
Llama 3.2 11B	19.3	30.7	13.8	30.8	41.2	38.1	29.4
Gemma 3 12B	27.3	38.5	23.7	38.4	44.3	38.0	30.4
Qwen3.5 9B	29.6	38.7	26.3	40.0	46.1	42.3	32.3
Aya Vision 8B	27.8	39.1	4.5	39.8	47.3	45.6	34.8
InternVL 3.5 8B	17.3	43.9	11.5	31.7	41.7	39.0	28.9
GLM-4.6V Flash	25.1	44.2	13.7	31.1	46.8	42.0	31.2
GPT-4.1	33.7	41.4	26.1	37.4	43.6	40.3	32.7
GPT-5 mini	**41.7**	**51.1**	**36.7**	**46.3**	**52.3**	**49.2**	**42.5**
o4-mini	37.1	46.6	32.1	43.5	49.3	42.7	37.5
with slide text							
Llama 4 Scout	27.4	36.8	24.7	40.0	49.3	**47.0**	37.3
Llama 3.2 11B	17.7	29.7	13.9	30.8	42.0	38.8	28.1
Gemma 3 12B	29.7	40.6	26.4	42.7	49.7	44.4	34.9
Qwen3.5 9B	29.7	38.6	26.3	39.9	45.6	42.6	32.3
Aya Vision 8B	25.6	40.1	3.0	39.3	46.7	43.8	33.4
InternVL 3.5 8B	18.5	42.9	12.4	31.3	41.9	39.8	29.5
GLM-4.6V Flash	25.6	45.1	14.4	30.9	46.0	41.0	28.2
GPT-4.1	30.9	37.7	22.9	32.9	38.0	35.4	28.2
GPT-5 mini	**39.6**	**48.6**	**34.0**	**44.9**	**50.0**	45.1	**39.9**
o4-mini	33.2	41.8	27.9	41.5	48.3	40.7	37.0
with slide image							
Llama 4 Scout	28.7	39.9	24.9	41.1	50.1	46.6	37.8
Llama 3.2 11B	3.8	13.2	4.4	23.8	37.4	31.6	22.3
Gemma 3 12B	27.6	39.7	23.9	41.3	47.5	41.3	32.8
Qwen3.5 9B	30.9	43.7	28.3	43.8	50.0	46.4	37.1
Aya Vision 8B	26.2	39.9	2.6	39.5	46.7	45.2	33.8
InternVL 3.5 8B	20.1	43.5	11.6	33.5	44.0	39.2	28.3
GLM-4.6V Flash	22.9	41.9	15.1	27.9	42.1	37.2	28.2
GPT-4.1	35.3	43.5	27.1	40.4	46.9	43.6	35.8
GPT-5 mini	**41.7**	**51.1**	**36.3**	**46.8**	**51.6**	**48.2**	**41.6**
o4-mini	36.6	46.0	31.3	42.8	48.7	40.7	37.0

Table 15: Average BLEU score per language in Translation of Aligned ASR Transcripts.

Transcript / Transl.	JA	ZH	TH	ID	ES	FR	DE
			KIWI (EN → ∗)				
no additional context							
Llama 4 Scout	79.9	74.7	74.1	75.7	75.8	77.0	75.5
Llama 3.2 11B	77.1	73.1	70.0	74.3	74.9	76.1	73.1
Gemma 3 12B	79.9	75.7	75.7	77.4	77.2	78.1	75.5
Qwen3.5 9B	80.6	76.8	75.7	77.5	76.9	78.3	75.9
Aya Vision 8B	79.1	74.8	54.8	75.1	75.7	76.8	74.1
InternVL 3.5 8B	75.6	74.2	66.4	72.6	74.7	75.2	72.1
GLM-4.6V Flash	78.5	74.7	69.0	71.7	75.0	75.2	72.8
GPT-4.1	81.2	76.6	76.2	**78.6**	78.5	79.2	76.8
GPT-5 mini	81.1	76.7	76.2	78.0	**78.8**	**79.5**	**77.1**
o4-mini	**81.3**	**76.8**	**76.3**	78.3	78.6	79.4	76.9
with slide text							
Llama 4 Scout	79.6	74.6	74.1	75.9	76.0	77.0	75.7
Llama 3.2 11B	75.7	72.4	69.5	74.1	74.8	75.9	72.8
Gemma 3 12B	79.8	75.6	75.3	77.1	76.8	78.2	75.8
Qwen3.5 9B	80.7	76.9	75.8	77.7	77.1	78.8	76.2
Aya Vision 8B	79.1	74.7	53.0	75.4	75.7	77.0	74.4
InternVL 3.5 8B	75.9	74.4	65.7	72.9	74.7	75.3	72.7
GLM-4.6V Flash	78.8	74.7	69.3	71.9	75.1	75.7	72.8
GPT-4.1	81.0	76.3	75.7	78.2	78.3	79.0	76.5
GPT-5 mini	**81.4**	**76.9**	**76.1**	**78.3**	**78.7**	**79.5**	**77.0**
o4-mini	79.7	75.0	74.3	76.5	77.0	78.0	75.4
with slide image							
Llama 4 Scout	79.8	75.1	73.8	75.3	75.6	76.8	75.6
Llama 3.2 11B	65.1	60.5	61.4	70.7	73.1	73.0	69.7
Gemma 3 12B	79.6	75.7	75.2	77.5	77.1	78.3	75.9
Qwen3.5 9B	80.5	76.6	75.4	76.8	76.7	78.4	75.9
Aya Vision 8B	79.2	74.4	52.8	74.9	75.8	77.0	74.1
InternVL 3.5 8B	76.3	74.9	65.5	73.5	75.1	75.8	72.5
GLM-4.6V Flash	77.5	74.4	67.9	71.5	74.6	72.6	72.0
GPT-4.1	81.2	76.8	76.3	**78.6**	78.5	79.5	**77.0**
GPT-5 mini	**81.3**	76.8	76.3	78.4	**78.7**	**79.6**	77.0
o4-mini	81.3	**76.8**	**76.3**	78.3	78.6	79.5	76.9

Table 17: Average COMET-KIWI score per language in Translation of Aligned ASR Transcripts.

Transcript / Transl.	JA	ZH	TH	ID	ES	FR	DE
			COMET (EN → ∗)				
no additional context							
Llama 4 Scout	86.8	82.8	86.5	86.3	82.3	80.9	81.7
Llama 3.2 11B	83.4	81.0	82.8	84.7	80.9	79.0	78.5
Gemma 3 12B	87.1	83.8	87.8	87.5	83.3	81.7	82.0
Qwen3.5 9B	87.7	84.7	88.0	87.7	83.2	82.3	82.3
Aya Vision 8B	85.9	82.3	69.5	85.8	82.0	80.5	80.0
InternVL 3.5 8B	81.7	82.5	76.0	83.5	80.7	78.4	76.9
GLM-4.6V Flash	85.4	83.2	82.0	82.5	81.6	78.6	78.5
GPT-4.1	88.9	85.7	89.0	88.9	85.2	83.9	84.2
GPT-5 mini	**89.7**	**86.7**	**89.7**	**89.6**	**86.4**	**84.8**	**85.3**
o4-mini	89.6	86.6	89.5	89.6	86.0	84.5	84.8
with slide text							
Llama 4 Scout	86.7	82.6	86.3	86.4	82.4	81.1	82.1
Llama 3.2 11B	82.8	80.7	82.3	84.5	80.7	78.8	77.9
Gemma 3 12B	87.4	84.0	87.5	87.7	83.6	82.1	82.3
Qwen3.5 9B	87.9	85.0	88.0	87.9	83.3	82.4	82.5
Aya Vision 8B	86.1	82.5	67.3	85.8	81.9	80.4	79.9
InternVL 3.5 8B	81.8	83.0	75.9	83.6	80.6	78.7	77.6
GLM-4.6V Flash	85.5	83.9	82.3	83.1	81.7	79.0	78.5
GPT-4.1	88.6	85.5	88.5	88.4	84.6	83.1	83.5
GPT-5 mini	**89.5**	**86.5**	**89.3**	**89.5**	**86.0**	**84.3**	**84.9**
o4-mini	87.4	83.8	86.6	87.2	84.1	82.3	82.6
with slide image							
Llama 4 Scout	86.8	83.2	86.1	86.1	82.2	80.5	81.8
Llama 3.2 11B	69.1	68.4	74.9	82.1	79.0	75.2	75.0
Gemma 3 12B	87.2	84.2	87.5	87.9	83.7	82.0	82.1
Qwen3.5 9B	87.6	85.2	87.7	87.6	83.1	82.3	82.2
Aya Vision 8B	86.0	82.5	66.5	86.0	82.0	80.8	79.9
InternVL 3.5 8B	82.1	83.5	74.6	84.2	81.2	79.0	77.6
GLM-4.6V Flash	84.2	80.7	71.9	82.6	80.7	76.5	76.3
GPT-4.1	88.9	86.1	89.1	89.2	85.6	84.3	84.4
GPT-5 mini	**89.8**	**86.9**	**89.8**	**89.7**	**86.2**	**84.9**	**85.3**
o4-mini	89.5	86.3	89.4	89.5	86.0	84.2	84.7

Table 16: Average COMET score per language in Translation of Aligned ASR Transcripts.

Transcript / Transl.	JA	ZH	TH	ID	ES	FR	DE
			NIST (EN → ∗)				
no additional context							
Llama 4 Scout	5.1	6.1	5.1	6.2	7.0	6.5	6.0
Llama 3.2 11B	4.3	5.7	3.7	5.5	6.3	6.0	5.3
Gemma 3 12B	5.1	6.4	5.1	6.2	6.5	6.1	5.5
Qwen3.5 9B	5.3	6.4	5.3	6.2	6.6	6.4	5.6
Aya Vision 8B	5.1	6.5	1.8	6.2	6.7	6.6	5.8
InternVL 3.5 8B	3.3	6.9	2.9	5.4	6.1	5.9	5.2
GLM-4.6V Flash	4.7	6.9	3.3	5.5	6.8	6.4	5.4
GPT-4.1	5.7	6.7	5.2	6.2	6.5	6.2	5.7
GPT-5 mini	**6.4**	**7.5**	**6.3**	**6.8**	**7.3**	**7.0**	**6.6**
o4-mini	5.9	7.1	5.8	6.5	7.0	6.4	6.0
with slide text							
Llama 4 Scout	4.9	5.9	5.0	6.2	6.8	6.6	5.9
Llama 3.2 11B	3.9	5.5	3.9	5.5	6.3	6.1	5.2
Gemma 3 12B	5.3	6.6	5.3	6.5	7.0	6.6	5.9
Qwen3.5 9B	5.3	6.4	5.3	6.2	6.6	6.4	5.7
Aya Vision 8B	4.7	6.5	1.2	6.2	6.7	6.4	5.7
InternVL 3.5 8B	3.6	6.7	3.1	5.3	6.1	6.1	5.3
GLM-4.6V Flash	4.8	6.9	3.5	5.4	6.8	6.2	4.8
GPT-4.1	5.4	6.2	4.8	5.6	5.9	5.7	5.1
GPT-5 mini	**6.3**	**7.3**	**6.0**	**6.7**	**7.1**	**6.7**	**6.3**
o4-mini	5.4	6.3	5.0	6.1	6.6	6.0	5.7
with slide image							
Llama 4 Scout	5.2	6.4	5.1	6.3	7.0	6.6	6.0
Llama 3.2 11B	1.0	2.5	1.5	4.6	5.9	5.4	4.5
Gemma 3 12B	5.1	6.5	4.9	6.4	6.8	6.3	5.7
Qwen3.5 9B	5.4	6.9	5.5	6.5	6.9	6.7	6.0
Aya Vision 8B	4.8	6.6	1.1	6.3	6.7	6.7	5.7
InternVL 3.5 8B	4.0	6.9	2.9	5.7	6.5	6.0	5.1
GLM-4.6V Flash	4.3	6.4	3.7	5.0	6.2	5.8	4.9
GPT-4.1	5.9	6.9	5.4	6.5	6.9	6.6	6.0
GPT-5 mini	**6.4**	**7.5**	**6.2**	**6.9**	**7.2**	**6.9**	**6.5**
o4-mini	5.9	7.0	5.7	6.5	6.9	6.2	6.0

Table 18: Average NIST score per language in Translation of Aligned ASR Transcripts.

Model	Slide Translation (EN → ∗)													
	no-context							*with slide image*						
	JA	ZH	TH	ID	ES	FR	DE	JA	ZH	TH	ID	ES	FR	DE
							ChrF							
Llama 3.2 11B	34.0	39.1	50.1	69.0	77.0	73.1	66.7	8.9	13.3	21.5	37.1	61.0	54.1	52.0
Gemma 3 12B	49.3	51.0	68.5	82.2	83.8	80.5	78.3	48.9	50.4	68.9	82.5	85.3	82.2	79.3
GPT-4.1	53.9	57.3	72.5	83.7	87.0	**85.0**	83.0	56.0	55.9	73.2	**85.2**	**87.7**	84.9	83.6
GPT-5 mini	**56.5**	**59.7**	**74.0**	**84.8**	**87.2**	84.2	**83.1**	**57.7**	**60.3**	**74.2**	85.1	87.4	**85.4**	**83.8**
o4-mini	55.2	57.9	73.7	83.3	86.5	83.7	83.0	55.2	58.5	73.7	84.5	86.7	84.1	83.2
							BLEU							
Llama 3.2 11B	14.7	27.5	15.8	25.0	35.1	25.0	14.7	2.1	5.3	2.4	6.7	22.2	13.6	13.9
Gemma 3 12B	43.0	55.5	45.8	57.8	52.9	46.7	46.3	30.1	40.5	33.9	48.4	51.5	46.1	44.8
GPT-4.1	48.6	63.1	52.8	70.3	75.1	**69.4**	66.8	51.1	62.1	53.6	**72.1**	**76.4**	69.4	68.7
GPT-5 mini	**51.2**	**65.8**	**54.1**	**70.8**	**75.2**	68.3	67.1	**52.7**	**66.3**	**54.2**	71.8	76.0	**70.4**	**69.2**
o4-mini	49.9	64.1	52.1	70.1	74.7	67.7	**68.2**	49.8	65.1	53.0	71.9	74.8	68.6	68.5
							COMET							
Llama 3.2 11B	70.4	64.6	69.4	78.9	74.4	73.9	73.2	65.2	65.2	63.9	72.4	76.7	70.4	74.6
Gemma 3 12B	82.9	85.8	82.6	86.5	77.5	78.0	79.6	83.7	85.5	84.5	88.0	87.3	87.1	86.4
GPT-4.1	89.7	92.5	91.8	94.1	**93.8**	**93.7**	**93.6**	89.6	91.8	92.2	**94.8**	**93.9**	**93.6**	93.5
GPT-5 mini	**90.2**	**93.0**	**92.0**	**94.5**	93.5	93.4	93.5	89.8	**92.9**	**92.3**	94.7	93.8	93.6	**93.7**
o4-mini	90.0	92.8	92.0	93.8	93.3	93.5	93.5	**90.0**	92.7	91.8	94.6	93.6	93.5	93.4
							COMET-KIWI							
Llama 3.2 11B	69.8	71.1	68.9	75.5	77.4	77.2	75.2	61.5	62.6	57.9	67.9	73.7	71.9	72.1
Gemma 3 12B	72.7	77.6	74.6	77.7	78.5	78.7	77.4	73.1	76.5	74.3	76.9	79.2	78.9	78.2
GPT-4.1	75.3	80.2	78.2	78.7	82.2	82.2	80.9	75.1	78.4	78.1	80.2	82.3	**82.2**	80.7
GPT-5 mini	75.4	**80.5**	**78.7**	**80.3**	82.4	82.1	81.0	75.2	**80.4**	**78.5**	80.2	82.5	82.2	81.0
o4-mini	**75.4**	80.3	78.3	79.1	**82.4**	**82.3**	**81.1**	**75.4**	80.4	78.1	**80.5**	**82.5**	82.1	**81.1**
							NIST							
Llama 3.2 11B	2.6	3.6	2.7	3.1	4.0	3.0	1.9	0.4	0.9	0.5	1.0	2.7	1.8	1.8
Gemma 3 12B	6.1	6.6	5.7	6.0	5.5	5.2	5.0	4.6	5.0	4.4	5.2	5.4	5.1	5.0
GPT-4.1	6.8	7.3	6.3	7.3	**7.6**	**7.3**	6.9	6.9	7.3	6.4	**7.3**	**7.7**	**7.3**	7.1
GPT-5 mini	**7.0**	**7.4**	**6.4**	7.3	7.6	7.2	7.0	**7.1**	**7.5**	**6.4**	7.3	7.7	7.3	**7.1**
o4-mini	6.9	7.3	6.3	**7.3**	7.6	7.2	**7.1**	6.9	7.4	6.3	7.3	7.6	7.2	7.1

Table 19: Average scores per language using various metrics in LECTRANS slide translation task.

Transcript Transl.	ChrF score per subject																
	anat-1	anat-2	bio-1	bio-2	bio-3	bio-4	dental	i-cs-prog	i-micro econs	i-neural -comp	i-physics	ml-1	psy-1	psy-2	speaking	theory -comp	urban -policy
								no additional context									
Llama 4 Scout	51.8	50.7	60.2	59.4	58.5	60.2	53.1	62.6	45.4	60.7	59.7	56.4	59.1	55.8	**57.3**	55.1	58.8
Llama 3.2 11B	46.4	42.9	50.5	52.1	52.7	53.5	45.2	54.0	39.9	52.2	51.8	50.2	52.2	49.7	46.4	48.8	50.5
Gemma 3 12B	49.6	48.5	56.2	55.7	59.0	58.1	50.2	58.1	44.1	57.6	56.6	52.6	57.3	52.1	52.0	51.9	55.7
Qwen3.5 9B	52.6	48.9	59.3	58.8	60.1	59.1	53.0	60.2	46.8	60.3	58.5	54.2	58.6	53.9	51.2	54.3	57.4
Aya Vision 8B	46.5	45.1	53.5	55.1	52.9	55.8	46.9	58.1	42.0	53.5	53.8	53.3	54.6	52.5	51.9	51.2	53.8
InternVL 3.5 8B	46.0	44.5	55.0	52.9	52.8	55.4	45.3	56.9	41.3	51.9	50.4	48.1	52.8	45.5	47.0	51.8	50.7
GLM-4.6V Flash	49.1	46.4	55.6	54.5	54.9	57.8	46.9	58.5	43.9	53.6	54.9	53.1	55.7	53.9	50.7	50.3	54.1
GPT-4.1	53.7	49.4	57.7	57.5	61.4	59.9	53.5	61.7	45.6	60.3	58.1	54.7	59.9	52.7	49.4	55.1	58.1
GPT-5 mini	**59.3**	**55.4**	**62.3**	**66.6**	**66.2**	**64.7**	**62.6**	**70.1**	**51.1**	**66.3**	**65.9**	**59.5**	**70.1**	**61.7**	53.7	**61.1**	**64.8**
o4-mini	56.3	53.8	61.0	63.4	64.0	61.4	58.6	67.4	48.9	63.6	61.7	53.3	66.2	57.3	48.4	58.9	62.2
								with slide text									
Llama 4 Scout	52.1	51.3	57.1	58.9	59.2	59.8	54.0	62.2	46.1	61.1	58.8	56.7	58.9	56.1	**57.7**	55.8	58.6
Llama 3.2 11B	46.8	41.7	52.4	51.0	52.2	52.6	44.6	53.3	40.2	51.6	51.5	49.9	51.9	48.8	47.2	49.2	50.3
Gemma 3 12B	52.8	48.4	**61.1**	56.7	61.9	61.7	51.7	61.7	47.2	59.5	60.1	57.2	58.8	53.1	57.2	54.6	59.6
Qwen3.5 9B	54.3	49.1	58.9	57.6	61.5	58.7	53.5	60.3	47.3	60.1	58.0	54.3	57.3	53.0	52.5	54.5	58.4
Aya Vision 8B	47.7	45.0	51.3	54.3	53.4	54.1	44.8	58.6	41.1	52.6	54.1	50.2	53.4	51.4	51.5	51.2	53.6
InternVL 3.5 8B	47.4	44.6	50.8	52.9	54.3	54.5	45.9	57.3	42.1	52.6	53.3	47.7	52.4	45.5	46.7	51.8	53.9
GLM-4.6V Flash	49.2	46.1	56.4	53.2	51.7	57.9	45.7	57.9	43.2	56.6	56.2	54.5	56.1	53.3	50.4	50.1	52.5
GPT-4.1	51.4	47.3	48.8	54.5	59.6	57.5	52.4	59.1	38.7	58.9	55.6	51.7	58.1	50.3	45.8	52.8	54.0
GPT-5 mini	**59.2**	**55.3**	58.3	**64.8**	**62.2**	**63.4**	**59.6**	**69.2**	**48.6**	**63.7**	**63.6**	**59.0**	**66.4**	**58.3**	52.1	**58.8**	**64.1**
o4-mini	52.6	52.4	57.0	63.9	54.5	54.5	56.6	65.1	46.1	63.3	58.6	57.0	62.7	55.9	43.2	56.1	59.8
								with slide image									
Llama 4 Scout	53.2	50.3	59.0	59.4	59.3	60.6	53.8	62.6	45.9	61.5	59.9	57.0	59.1	55.8	**58.3**	55.3	58.9
Llama 3.2 11B	38.2	33.3	42.7	42.5	40.4	46.3	36.4	43.6	30.7	40.0	41.8	39.5	40.2	41.5	37.4	40.0	44.0
Gemma 3 12B	50.0	48.0	58.0	56.5	60.5	60.0	52.1	60.5	45.9	57.8	58.0	54.3	58.3	51.9	55.7	54.4	57.5
Qwen3.5 9B	56.1	50.3	61.3	60.4	62.0	60.6	55.0	62.5	48.4	61.8	59.9	57.1	60.7	55.9	57.2	57.1	60.3
Aya Vision 8B	45.7	45.5	52.9	53.6	53.0	53.2	45.3	58.2	41.2	53.1	53.9	51.9	53.6	51.7	49.7	51.0	52.3
InternVL 3.5 8B	46.2	44.2	54.0	52.3	53.2	54.6	45.2	56.2	42.3	52.9	53.0	49.0	52.4	46.2	47.9	51.3	53.9
GLM-4.6V Flash	46.1	44.7	51.5	54.2	53.8	56.2	47.1	56.6	41.9	48.5	56.0	54.1	53.4	51.1	53.3	49.6	54.0
GPT-4.1	54.7	50.8	59.2	59.4	62.1	61.5	55.1	63.1	48.2	62.5	59.7	58.1	62.3	55.5	50.9	56.6	60.0
GPT-5 mini	**60.0**	**55.1**	**63.0**	**67.2**	**65.8**	**63.9**	**62.4**	**71.2**	**50.4**	**64.9**	**65.6**	**60.9**	**68.5**	**56.9**	56.0	**60.8**	**65.1**
o4-mini	54.0	53.9	59.6	62.4	63.7	63.2	58.6	66.7	48.2	63.9	61.2	54.0	63.4	54.8	47.4	58.6	62.5

Table 20: Average ChrF score in Translation Aligned ASR Transcript per Lecture Subject

Transcript Transl.	BLEU score per subject																
	anat-1	anat-2	bio-1	bio-2	bio-3	bio-4	dental	i-cs-prog	i-micro econs	i-neural -comp	i-physics	ml-1	psy-1	psy-2	speaking	theory -comp	urban -policy
no additional context																	
Llama 4 Scout	33.7	32.5	40.5	39.9	39.5	40.5	34.3	45.9	27.5	43.7	41.8	35.9	38.0	36.4	**36.3**	37.5	39.9
Llama 3.2 11B	27.4	23.5	29.9	32.4	32.5	33.2	24.5	36.1	20.0	33.1	31.1	27.7	31.1	27.8	23.4	29.5	30.6
Gemma 3 12B	30.8	29.4	36.2	36.5	40.4	39.4	30.0	40.5	25.6	39.7	37.0	30.7	38.2	30.6	29.5	33.0	36.7
Qwen3.5 9B	35.3	29.8	39.3	39.8	42.3	40.8	33.6	42.9	28.4	42.7	39.5	31.7	38.4	32.8	27.7	36.5	38.8
Aya Vision 8B	30.3	28.6	34.9	38.1	35.6	38.9	28.9	42.0	24.7	37.3	36.9	33.3	36.2	32.8	31.5	33.9	36.3
InternVL 3.5 8B	24.5	27.1	35.0	34.2	32.8	36.0	25.1	40.1	22.7	32.0	32.1	29.0	30.5	28.3	26.1	33.9	30.2
GLM-4.6V Flash	32.0	28.2	36.4	34.9	35.3	38.8	27.2	41.3	24.7	38.1	35.5	31.2	35.4	33.5	29.6	31.4	34.6
GPT-4.1	35.4	29.8	37.6	39.8	43.9	40.8	33.4	44.6	26.2	41.9	38.5	34.2	41.1	31.4	26.4	36.3	38.7
GPT-5 mini	**42.7**	**37.7**	**44.5**	**51.7**	**51.0**	**49.0**	**46.8**	**56.5**	**33.2**	**49.7**	**49.5**	**41.1**	**55.5**	**44.2**	31.7	**43.2**	**48.8**
o4-mini	38.3	35.9	42.8	47.7	46.9	44.5	40.9	53.2	29.6	46.1	43.0	33.6	49.8	38.1	25.1	40.9	45.2
with slide text																	
Llama 4 Scout	32.2	33.5	37.0	39.5	38.7	40.3	35.2	45.1	27.9	43.9	40.3	35.9	37.0	36.4	36.5	37.9	39.9
Llama 3.2 11B	27.0	22.2	31.3	31.2	31.6	32.7	23.7	34.5	20.4	32.0	31.3	27.2	30.7	27.4	23.8	30.2	30.7
Gemma 3 12B	36.0	29.5	**42.5**	38.6	44.9	44.6	31.8	45.3	28.8	42.2	42.5	36.7	40.3	31.7	**36.7**	37.0	42.5
Qwen3.5 9B	36.6	30.1	38.4	38.6	43.5	40.4	33.8	42.9	29.0	42.4	38.7	32.0	36.5	31.3	29.5	36.1	39.6
Aya Vision 8B	30.8	28.4	30.3	37.3	35.3	34.9	26.4	41.8	24.8	36.5	37.2	30.3	36.3	32.0	30.8	33.8	35.8
InternVL 3.5 8B	27.1	26.0	25.8	35.0	34.6	34.9	25.1	40.8	24.2	34.6	35.9	28.4	31.3	26.9	26.6	33.7	34.8
GLM-4.6V Flash	31.7	27.9	36.5	33.4	32.1	39.2	26.3	38.8	24.8	38.0	37.4	32.0	35.5	32.5	29.9	31.1	34.2
GPT-4.1	32.1	27.4	23.6	35.6	40.8	37.6	32.3	41.0	16.5	39.6	35.4	31.1	38.4	28.6	22.8	33.0	33.2
GPT-5 mini	**41.8**	**37.7**	40.3	**49.4**	**45.9**	**47.0**	**42.6**	**55.4**	**30.9**	45.9	**46.1**	**40.9**	**50.1**	**39.5**	30.3	**41.4**	**48.2**
o4-mini	34.8	33.5	38.4	48.2	35.9	37.9	39.4	50.4	27.5	**46.4**	40.4	38.1	45.9	36.0	22.8	38.3	42.9
with slide image																	
Llama 4 Scout	36.0	33.1	38.7	40.9	40.8	42.0	35.0	46.0	28.1	44.5	41.8	36.8	37.8	36.0	**37.3**	38.0	40.8
Llama 3.2 11B	18.7	15.1	21.5	23.3	18.0	26.6	14.8	23.7	12.6	20.1	20.9	19.0	18.5	20.2	15.3	20.3	23.0
Gemma 3 12B	31.8	28.5	38.7	38.2	42.9	42.4	32.3	43.5	27.4	39.2	39.3	32.6	39.5	31.0	34.3	36.4	39.1
Qwen3.5 9B	39.8	32.7	42.9	43.0	45.2	43.3	36.6	46.1	31.1	45.6	42.2	36.2	41.2	35.4	36.5	39.9	42.8
Aya Vision 8B	28.6	28.9	34.7	37.7	35.6	36.8	26.7	42.2	24.1	36.6	36.6	33.2	35.7	31.8	29.7	33.8	35.2
InternVL 3.5 8B	29.0	25.6	33.4	32.9	33.2	35.2	25.9	38.8	24.3	34.9	34.6	29.6	32.2	28.2	27.9	33.5	35.3
GLM-4.6V Flash	24.4	27.3	29.5	33.3	32.6	36.4	27.8	36.7	23.8	27.9	36.9	32.0	30.5	28.9	30.1	30.2	34.5
GPT-4.1	37.2	32.3	39.8	41.9	44.4	43.3	35.7	46.9	28.9	44.8	41.1	38.8	44.8	35.7	27.6	38.2	41.1
GPT-5 mini	**42.5**	**37.5**	**46.1**	**52.5**	**50.0**	**46.8**	**46.3**	**58.1**	**32.8**	**47.8**	**48.4**	**42.5**	**53.7**	**38.3**	34.3	**44.0**	**49.0**
o4-mini	35.3	35.3	40.7	46.0	46.8	45.7	40.8	52.1	29.3	45.7	42.9	33.9	46.4	35.1	25.2	40.7	45.4

Table 21: Average BLEU score in Translation Aligned ASR Transcript per Lecture Subject

Transcript Transl.	COMET score per subject																
	anat-1	anat-2	bio-1	bio-2	bio-3	bio-4	dental	i-cs-prog	i-micro econs	i-neural -comp	i-physics	ml-1	psy-1	psy-2	speaking	theory -comp	urban -policy
no additional context																	
Llama 4 Scout	79.2	80.1	85.5	89.1	85.2	87.8	81.1	86.8	73.5	83.4	86.1	81.7	87.5	85.9	86.6	81.8	85.0
Llama 3.2 11B	77.6	77.3	83.0	86.9	83.3	86.3	77.4	84.8	70.0	80.9	83.7	78.9	84.5	84.4	84.3	79.1	82.6
Gemma 3 12B	80.8	81.5	85.9	89.5	87.9	88.9	81.7	87.4	74.7	83.5	87.0	82.3	87.8	86.5	87.2	82.2	85.9
Qwen3.5 9B	81.6	80.9	86.7	89.9	88.0	88.9	82.7	87.3	75.7	84.3	87.4	82.3	88.2	87.2	87.5	82.7	86.0
Aya Vision 8B	76.4	76.4	82.1	85.8	81.2	85.8	76.7	84.5	70.6	79.3	83.0	79.7	84.3	84.2	84.1	78.7	81.9
InternVL 3.5 8B	75.6	75.3	82.6	85.2	82.3	84.5	76.0	84.0	68.8	79.5	81.8	77.7	83.8	80.1	82.0	78.3	81.3
GLM-4.6V Flash	77.8	77.1	83.7	86.5	83.3	86.5	77.4	85.0	71.1	80.9	84.4	79.8	85.2	84.3	84.4	79.1	82.4
GPT-4.1	83.6	83.1	87.2	90.4	89.1	90.2	84.5	89.1	76.6	85.5	88.4	85.6	90.0	87.7	88.0	84.2	87.5
GPT-5 mini	**84.3**	**83.5**	88.1	**91.5**	**89.9**	**91.1**	**85.6**	**90.5**	77.5	**86.2**	**89.5**	**86.9**	91.0	**89.3**	**88.9**	**85.2**	87.7
o4-mini	84.0	83.4	**88.1**	91.1	89.8	90.6	85.2	90.0	**77.8**	86.1	89.1	86.6	**91.4**	88.8	88.4	84.5	**87.7**
with slide text																	
Llama 4 Scout	79.9	81.1	85.5	88.7	85.3	87.8	81.5	86.3	72.9	83.4	86.0	82.1	87.1	86.4	85.9	81.9	85.1
Llama 3.2 11B	77.4	77.3	82.5	86.1	82.3	86.6	77.3	84.5	69.9	80.9	83.6	78.2	84.0	83.5	83.6	78.9	82.5
Gemma 3 12B	81.6	81.6	86.8	89.4	87.5	89.1	82.1	87.3	74.7	84.1	86.8	82.8	87.9	86.1	87.6	82.2	86.3
Qwen3.5 9B	82.7	81.5	86.7	89.7	88.3	88.7	82.9	87.2	76.4	84.7	87.3	82.7	87.9	86.4	87.3	82.8	86.3
Aya Vision 8B	76.1	76.5	81.3	85.5	81.7	85.1	76.2	84.5	70.2	78.6	83.2	78.0	84.1	83.9	83.9	78.7	82.0
InternVL 3.5 8B	76.1	76.0	82.5	85.4	82.4	85.1	76.3	84.2	69.5	79.6	82.8	77.3	83.9	80.4	81.3	78.2	82.1
GLM-4.6V Flash	79.0	77.9	83.5	86.6	83.3	86.4	77.9	85.5	71.6	81.5	84.8	80.1	85.2	84.4	84.4	79.4	82.9
GPT-4.1	83.3	82.6	86.7	89.8	**89.6**	90.1	83.9	88.7	76.1	85.4	87.5	85.0	89.4	86.9	87.5	83.6	86.7
GPT-5 mini	**85.0**	**84.2**	**87.1**	**91.4**	89.1	**90.5**	**84.9**	**90.3**	**77.5**	**86.2**	**89.2**	**86.4**	**90.9**	**88.3**	**88.4**	**84.3**	**88.0**
o4-mini	82.9	83.7	84.2	90.9	78.2	84.1	84.2	88.7	75.3	85.6	85.1	86.4	89.5	87.9	85.6	82.8	87.1
with slide image																	
Llama 4 Scout	80.0	79.9	85.7	88.8	85.3	87.8	81.0	86.5	72.5	83.2	86.0	82.0	87.4	85.9	86.3	81.4	85.3
Llama 3.2 11B	71.8	67.7	78.2	80.9	76.6	82.1	71.6	76.1	62.1	72.3	76.4	71.2	77.0	78.0	80.9	72.0	77.1
Gemma 3 12B	81.8	81.8	86.2	89.3	88.2	89.3	82.4	87.3	75.0	84.2	86.9	82.2	87.9	86.0	87.3	82.1	86.0
Qwen3.5 9B	82.3	80.9	86.7	89.6	87.6	88.7	82.6	87.3	75.3	84.4	87.1	82.7	88.2	87.0	87.8	82.5	85.9
Aya Vision 8B	75.6	76.4	82.0	85.3	81.6	84.7	76.3	84.7	70.2	79.0	82.9	78.5	84.1	84.0	83.7	78.6	81.5
InternVL 3.5 8B	76.3	76.4	82.6	85.4	82.8	84.9	76.1	84.0	70.0	79.8	82.9	78.3	83.6	79.8	81.9	78.0	82.5
GLM-4.6V Flash	75.7	73.5	80.5	81.7	79.9	83.2	74.2	81.9	68.7	77.5	82.8	79.3	80.6	80.9	82.6	77.4	82.4
GPT-4.1	84.4	83.1	87.7	90.8	89.2	90.3	84.3	89.2	76.7	85.6	88.5	86.2	90.4	88.4	88.4	84.6	87.4
GPT-5 mini	**85.0**	**84.1**	**88.0**	**91.6**	**90.1**	**91.0**	**85.6**	**91.0**	**77.7**	86.0	**89.2**	**87.3**	90.9	88.4	**88.7**	**85.4**	**88.0**
o4-mini	84.2	84.0	87.9	90.9	89.7	90.8	85.1	89.8	76.8	**86.3**	88.8	86.4	**91.0**	**88.5**	88.0	85.0	87.5

Table 22: Average COMET score in Translation Aligned ASR Transcript per Lecture Subject

Transcript Transl.	COMET-KIWI score per subject																
	anat-1	anat-2	bio-1	bio-2	bio-3	bio-4	dental	i-cs-prog	i-micro econs	i-neural -comp	i-physics	ml-1	psy-1	psy-2	speaking	theory -comp	urban -policy
no additional context																	
Llama 4 Scout	72.3	70.8	78.8	82.7	78.5	79.8	73.2	78.6	62.3	76.1	77.8	73.0	80.4	79.5	77.1	75.0	77.9
Llama 3.2 11B	70.9	68.2	77.2	80.6	77.3	78.3	70.5	76.8	59.4	74.2	75.8	70.3	78.1	77.7	75.7	72.5	75.7
Gemma 3 12B	73.3	71.1	79.4	83.4	81.2	81.1	73.9	79.2	63.3	76.8	79.2	74.1	80.9	80.1	77.9	76.2	79.0
Qwen3.5 9B	73.5	71.3	**80.5**	83.7	80.8	81.0	74.8	79.2	64.9	76.9	79.4	74.0	81.4	81.0	78.2	76.2	78.9
Aya Vision 8B	69.0	67.5	75.4	79.2	74.7	77.4	69.3	76.1	59.8	72.0	74.6	69.9	77.0	76.8	74.4	71.8	74.5
InternVL 3.5 8B	68.9	68.1	76.7	79.5	76.3	77.2	69.0	76.5	58.9	72.9	74.4	69.0	78.0	74.6	73.2	72.7	74.5
GLM-4.6V Flash	70.7	67.7	76.8	79.7	76.9	78.3	69.7	76.8	60.4	73.5	76.2	70.7	78.1	77.4	74.8	72.2	75.3
GPT-4.1	75.2	**72.0**	80.3	**84.0**	82.1	**81.4**	75.1	79.9	64.9	77.4	79.8	**78.1**	82.7	80.9	78.2	**77.1**	79.6
GPT-5 mini	75.3	71.8	79.8	84.0	**82.4**	81.0	**75.7**	**80.1**	65.6	77.7	**80.1**	77.9	82.5	80.9	78.3	76.9	79.3
o4-mini	**75.4**	71.3	80.4	84.0	82.3	80.8	75.2	80.0	**65.8**	**77.8**	80.0	77.9	**83.2**	**81.1**	**78.3**	76.7	**79.8**
with slide text																	
Llama 4 Scout	72.9	71.4	79.0	82.8	78.6	79.7	73.4	78.1	62.1	76.2	78.2	72.8	80.1	79.8	76.2	75.0	78.0
Llama 3.2 11B	70.9	67.6	76.4	80.5	77.0	78.7	69.9	76.2	59.7	73.5	75.6	69.1	77.5	77.1	74.3	72.1	75.1
Gemma 3 12B	74.0	71.9	79.9	83.4	80.9	80.8	73.7	79.3	63.3	76.7	79.0	73.5	80.9	79.4	77.6	75.9	78.4
Qwen3.5 9B	75.2	71.7	80.1	83.7	81.5	81.1	74.9	79.4	65.3	77.2	79.7	73.7	81.4	80.4	78.0	76.4	79.2
Aya Vision 8B	69.2	67.4	75.4	79.2	74.8	77.1	68.9	75.5	59.4	71.8	75.0	68.4	76.9	76.5	74.3	72.0	74.5
InternVL 3.5 8B	69.6	68.3	76.9	79.7	76.0	77.6	69.0	76.4	59.1	72.8	75.0	69.3	77.6	74.7	72.9	72.7	74.8
GLM-4.6V Flash	71.2	68.2	76.5	79.9	77.5	78.3	69.8	76.9	60.7	73.8	76.3	70.7	77.8	77.6	75.3	72.7	75.7
GPT-4.1	75.3	71.4	**80.3**	83.3	82.4	**81.5**	74.7	79.8	63.8	77.4	79.8	77.2	82.4	80.2	**78.1**	**76.9**	79.0
GPT-5 mini	**76.7**	72.4	79.2	**84.0**	**82.6**	81.0	**75.5**	**80.2**	**65.8**	**78.2**	**80.3**	77.5	**82.9**	80.3	78.0	76.3	**79.5**
o4-mini	75.2	**72.5**	77.0	83.9	74.1	76.1	75.3	78.6	63.2	77.5	76.7	**78.1**	81.8	**80.8**	77.1	74.5	79.2
with slide image																	
Llama 4 Scout	73.0	70.7	79.4	82.4	78.7	79.6	73.0	78.3	61.6	76.0	77.8	72.5	80.4	79.4	76.5	75.0	78.0
Llama 3.2 11B	64.4	60.6	71.5	75.1	70.0	74.7	64.5	69.0	53.4	66.2	68.6	62.8	70.5	71.2	72.3	65.1	70.1
Gemma 3 12B	74.4	71.3	80.0	83.5	81.6	81.0	74.3	79.0	63.4	76.9	78.9	73.7	80.4	79.4	77.4	75.7	78.8
Qwen3.5 9B	75.0	71.2	80.2	83.5	80.5	80.4	74.5	79.1	64.0	77.0	79.2	73.6	81.1	80.2	78.3	75.7	78.8
Aya Vision 8B	68.7	67.3	75.5	78.7	75.1	76.5	68.7	75.8	59.7	71.6	74.3	69.1	77.1	76.3	74.1	71.8	74.1
InternVL 3.5 8B	69.5	68.9	77.0	80.0	76.6	77.2	68.9	76.6	59.9	73.4	75.5	70.0	77.7	74.7	73.8	72.7	75.1
GLM-4.6V Flash	69.8	66.7	75.3	78.7	76.1	77.0	68.3	76.1	59.9	71.6	76.1	70.7	76.6	75.5	74.4	71.4	75.7
GPT-4.1	76.1	72.0	**80.4**	**84.3**	82.1	81.4	75.3	80.0	64.5	77.5	79.7	78.1	83.0	81.1	**79.0**	**76.8**	79.3
GPT-5 mini	**76.5**	72.2	79.8	83.9	**82.2**	81.2	**75.7**	**80.5**	**65.7**	77.5	80.0	**78.8**	82.8	80.5	78.3	76.4	79.3
o4-mini	76.4	**72.2**	79.9	83.6	82.2	**81.5**	75.7	79.8	64.8	**78.2**	**80.0**	78.0	**83.2**	**81.2**	77.7	76.5	**79.5**

Table 23: Average COMET-KIWI score in Translation Aligned ASR Transcript per Lecture Subject

Transcript Transl.	NIST score per subject																
	anat-1	anat-2	bio-1	bio-2	bio-3	bio-4	dental	i-cs-prog	i-micro econs	i-neural -comp	i-physics	ml-1	psy-1	psy-2	speaking	theory -comp	urban -policy
no additional context																	
Llama 4 Scout	5.3	5.2	6.3	6.2	6.1	5.9	5.9	6.5	4.8	6.5	6.6	6.1	6.0	6.1	**6.3**	5.9	6.3
Llama 3.2 11B	4.9	4.3	5.2	5.7	5.7	5.6	5.0	5.8	3.9	5.6	5.6	5.4	5.5	5.4	4.8	5.3	5.7
Gemma 3 12B	5.4	4.8	6.1	6.2	6.4	6.2	5.7	6.3	4.4	6.2	6.5	5.7	6.2	5.8	5.6	5.7	6.2
Qwen3.5 9B	5.8	4.9	6.3	6.4	6.5	6.2	5.9	6.4	4.7	6.5	6.6	5.8	6.2	5.9	5.5	5.9	6.3
Aya Vision 8B	5.0	4.7	5.6	6.0	5.5	5.8	5.2	6.1	4.3	5.7	6.0	5.6	5.8	5.8	5.8	5.4	6.0
InternVL 3.5 8B	4.3	4.6	5.7	5.5	5.3	5.4	4.8	5.9	4.1	5.2	5.4	5.0	5.2	5.1	4.6	5.6	4.9
GLM-4.6V Flash	5.4	4.7	5.9	5.5	5.8	6.0	5.1	6.2	4.3	6.1	5.9	5.5	5.9	5.9	5.5	5.1	5.7
GPT-4.1	5.8	4.5	6.4	6.6	6.7	6.3	6.0	6.8	4.2	6.5	6.6	6.0	6.7	5.9	5.1	5.9	6.5
GPT-5 mini	**6.5**	**5.3**	**6.9**	**7.7**	**7.3**	**6.8**	**7.2**	**7.8**	**5.1**	**7.2**	**7.6**	**6.5**	**7.8**	**7.0**	5.5	**6.6**	**7.4**
o4-mini	6.1	4.9	6.8	7.3	7.0	6.5	6.7	7.5	4.5	6.8	7.0	5.5	7.4	6.5	4.7	6.3	7.1
with slide text																	
Llama 4 Scout	5.1	5.3	5.7	6.1	5.9	5.8	6.0	6.3	4.7	6.5	6.3	6.1	5.8	6.1	6.3	6.0	6.2
Llama 3.2 11B	4.9	4.0	5.5	5.6	5.6	5.3	5.0	5.6	3.7	5.5	5.8	5.3	5.4	5.4	5.0	5.3	5.6
Gemma 3 12B	5.9	4.6	**6.6**	6.4	6.8	6.6	5.8	6.6	4.7	6.5	6.8	6.2	6.3	5.9	**6.4**	6.0	6.8
Qwen3.5 9B	5.9	5.0	6.1	6.2	6.6	6.1	5.9	6.4	4.9	6.4	6.5	5.8	6.0	5.7	5.8	5.9	6.4
Aya Vision 8B	5.0	4.7	4.9	5.8	5.5	5.1	4.7	6.1	4.2	5.5	5.9	5.3	5.6	5.6	5.7	5.5	5.8
InternVL 3.5 8B	4.7	4.3	4.5	5.7	5.7	5.3	4.8	6.2	4.2	5.5	6.0	4.8	5.2	4.9	4.8	5.6	6.0
GLM-4.6V Flash	5.2	4.6	6.0	5.4	5.2	6.0	4.9	5.9	4.4	6.0	6.4	5.8	5.9	5.9	5.5	5.0	5.6
GPT-4.1	5.5	4.3	4.0	6.2	6.5	6.0	5.8	6.4	2.9	6.3	6.3	5.7	6.4	5.5	4.4	5.5	6.0
GPT-5 mini	**6.6**	**5.3**	6.5	**7.4**	**6.9**	**6.8**	**6.8**	**7.7**	**5.0**	**6.9**	**7.3**	**6.5**	**7.4**	**6.6**	5.5	**6.4**	**7.3**
o4-mini	5.6	4.8	6.1	7.3	5.0	5.2	6.5	7.2	4.0	6.8	6.6	6.2	7.0	6.3	2.7	6.0	6.9
with slide image																	
Llama 4 Scout	5.6	5.2	6.0	6.3	6.2	6.1	6.1	6.5	4.8	6.6	6.7	6.2	6.0	6.1	**6.5**	6.0	6.5
Llama 3.2 11B	3.5	2.8	3.9	4.3	3.2	4.5	3.0	3.9	2.5	3.3	3.9	3.7	3.4	4.0	3.4	3.6	4.4
Gemma 3 12B	5.5	4.5	6.2	6.3	6.6	6.4	5.9	6.6	4.5	6.0	6.6	5.9	6.3	5.8	6.2	5.9	6.5
Qwen3.5 9B	6.1	**5.2**	6.5	6.6	6.7	6.4	6.2	6.7	**5.0**	6.7	6.8	6.2	6.3	6.1	6.4	6.2	6.7
Aya Vision 8B	4.6	4.8	5.5	5.8	5.6	5.5	4.8	6.1	4.1	5.6	6.0	5.5	5.6	5.6	5.5	5.5	5.7
InternVL 3.5 8B	4.9	4.3	5.6	5.4	5.4	5.4	4.9	5.8	4.3	5.6	6.0	5.1	5.5	5.1	5.2	5.5	6.0
GLM-4.6V Flash	4.1	4.6	4.8	5.5	5.4	5.6	5.1	5.6	4.3	4.2	6.3	5.8	5.1	5.2	5.8	4.8	5.8
GPT-4.1	6.0	4.8	6.6	6.8	6.8	6.6	6.2	7.0	4.6	6.8	6.9	6.5	7.0	6.3	5.3	6.1	6.8
GPT-5 mini	**6.5**	5.1	**7.0**	**7.7**	**7.2**	**6.7**	**7.1**	**7.9**	5.0	**7.0**	**7.5**	**6.8**	**7.6**	**6.4**	5.9	**6.6**	**7.4**
o4-mini	5.8	4.9	6.6	7.2	7.0	6.7	6.7	7.4	4.3	6.8	7.0	5.8	7.1	6.2	4.4	6.4	7.1

Table 24: Average NIST score in Translation Aligned ASR Transcript per Lecture Subject

Slide Transl.	anat-1	anat-2	bio-1	bio-2	bio-3	bio-4	dental	i-cs-prog	i-micro econs	i-neural -comp	i-physics	ml-1	psy-1	psy-2	speaking	theory -comp	urban -policy
Score per subject without additional context																	
ChrF																	
Llama 3.2 11B	54.0	60.1	60.7	61.8	60.7	55.0	43.9	61.7	49.3	55.5	61.9	74.0	56.5	58.7	53.7	66.4	59.3
Gemma 3 12B	65.0	71.0	76.0	77.3	71.1	61.0	52.7	72.2	62.7	74.2	75.1	83.8	68.5	72.2	67.5	76.4	72.3
GPT-4.1	69.8	70.9	79.7	**80.4**	73.0	63.4	59.8	78.6	**68.4**	80.0	**78.4**	83.2	74.7	76.9	72.9	**80.0**	78.8
GPT-5 mini	**71.1**	**75.1**	**80.4**	80.4	**75.8**	**65.0**	60.1	**80.2**	68.3	**81.1**	78.2	**87.0**	72.4	**79.4**	**73.3**	78.6	**80.0**
o4-mini	70.3	74.2	77.2	79.0	74.1	63.9	**60.7**	79.3	68.0	79.8	76.7	85.2	**74.8**	76.8	73.0	78.7	79.3
BLEU																	
Llama 3.2 11B	29.3	20.0	18.5	30.0	31.5	21.4	9.2	28.7	17.4	6.3	22.5	19.7	20.8	29.3	28.8	37.9	11.9
Gemma 3 12B	43.1	43.8	49.9	60.3	52.9	39.8	26.1	49.0	45.2	55.2	56.9	68.1	46.7	46.5	48.2	65.5	48.3
GPT-4.1	56.7	52.0	67.0	71.2	60.5	50.3	**42.9**	72.6	**60.8**	73.7	**69.8**	74.4	**63.1**	64.4	59.8	**75.5**	68.7
GPT-5 mini	**58.0**	**57.8**	**69.2**	**71.8**	**64.0**	**51.6**	39.6	**75.3**	59.6	73.8	68.7	**78.1**	60.4	**67.7**	61.2	71.8	**70.5**
o4-mini	57.9	56.2	65.2	70.4	62.6	51.3	42.5	73.1	58.9	**74.2**	66.7	76.7	63.0	63.4	**61.3**	72.3	69.4
COMET																	
Llama 3.2 11B	77.1	62.7	67.9	82.4	79.3	74.8	60.1	76.7	73.8	61.2	71.3	69.8	70.1	74.6	75.7	74.1	74.5
Gemma 3 12B	82.9	73.7	83.6	90.7	86.1	82.3	66.8	82.6	83.4	73.9	83.7	82.0	77.5	88.1	87.0	83.1	84.3
GPT-4.1	**90.5**	**93.0**	**94.3**	95.7	90.5	91.3	**86.3**	91.9	**91.9**	**94.6**	**95.7**	95.3	91.1	94.2	**93.4**	**91.7**	95.4
GPT-5 mini	90.0	92.3	94.3	**95.9**	92.1	**91.8**	86.0	**92.8**	91.7	94.3	95.4	**95.9**	91.2	**95.3**	93.2	91.0	**95.6**
o4-mini	90.1	92.3	92.4	95.7	**92.3**	90.4	85.7	92.4	91.8	94.4	95.2	95.7	**91.9**	94.9	93.3	91.7	95.6
COMET-KIWI																	
Llama 3.2 11B	79.1	70.5	72.5	83.3	76.8	73.2	63.5	77.7	70.9	66.9	71.0	68.6	73.1	76.4	77.9	74.3	75.2
Gemma 3 12B	82.4	73.3	77.6	85.4	80.1	74.7	64.1	79.9	74.5	68.7	76.1	73.4	75.0	81.6	81.4	77.6	79.2
GPT-4.1	**83.4**	79.5	**80.8**	87.0	77.9	**77.4**	69.7	81.6	75.4	78.0	80.6	77.8	79.4	82.1	81.8	**79.4**	82.7
GPT-5 mini	83.2	**80.3**	80.7	**87.0**	81.7	77.2	**70.5**	81.6	**75.6**	77.9	80.7	**78.4**	79.6	**82.4**	82.0	78.9	83.0
o4-mini	83.0	80.0	77.5	86.9	**81.8**	76.8	70.1	**81.8**	75.5	**78.0**	**80.8**	78.3	**79.8**	82.4	**82.1**	79.3	**83.0**
NIST																	
Llama 3.2 11B	4.5	2.5	2.7	3.3	4.9	3.6	1.4	3.3	2.6	0.7	2.9	2.0	2.6	3.8	4.3	4.2	1.6
Gemma 3 12B	5.7	4.9	6.0	5.6	7.2	6.0	4.0	5.2	6.2	5.2	6.3	6.3	5.2	5.5	6.1	6.6	5.2
GPT-4.1	6.9	5.9	7.7	6.4	8.3	7.1	**5.7**	7.3	**7.8**	7.0	**7.5**	7.4	**6.7**	7.0	7.2	**7.6**	6.9
GPT-5 mini	7.0	**6.2**	**7.8**	**6.4**	**8.4**	**7.2**	5.5	**7.5**	7.8	6.9	7.5	**7.6**	6.6	**7.2**	7.2	7.3	**7.0**
o4-mini	**7.0**	6.1	7.8	6.4	8.3	7.2	5.7	7.3	7.7	**7.0**	7.3	7.6	6.7	7.0	**7.2**	7.3	6.9

Table 25: Average scores per lecture subject using various metrics in LECTRANS slide translation task without any additional context.

Slide Transl.	anat-1	anat-2	bio-1	bio-2	bio-3	bio-4	dental	i-cs-prog	i-micro econs	i-neural -comp	i-physics	ml-1	psy-1	psy-2	speaking	theory -comp	urban -policy
Score per subject with full image provided																	
ChrF																	
Llama 3.2 11B	33.8	31.8	41.5	45.2	44.3	33.8	26.6	34.6	24.7	25.4	44.0	36.7	34.4	35.9	41.6	33.2	34.5
Gemma 3 12B	65.5	72.6	76.5	76.2	71.2	61.7	54.2	76.9	63.8	76.0	74.6	78.3	68.6	73.4	67.4	77.2	73.9
GPT-4.1	**71.8**	75.5	**80.3**	**81.0**	74.2	62.8	**61.2**	79.3	**68.4**	79.7	77.8	86.3	**75.0**	76.3	**73.6**	76.1	79.3
GPT-5 mini	71.3	**77.1**	80.3	80.6	**75.7**	**65.6**	60.6	**81.1**	68.3	**80.2**	**79.3**	**88.8**	74.1	**79.5**	73.2	**80.1**	**81.1**
o4-mini	70.3	74.9	78.9	80.5	74.0	63.4	61.2	80.1	67.4	79.4	77.4	87.0	73.7	78.4	73.3	77.9	79.4
BLEU																	
Llama 3.2 11B	8.9	4.2	7.6	17.8	17.0	7.6	2.8	10.6	5.1	1.8	15.0	6.5	7.9	11.3	18.9	12.5	5.2
Gemma 3 12B	47.4	28.3	41.3	55.1	48.2	42.2	17.5	40.2	49.0	33.3	23.6	49.8	37.8	56.2	49.2	45.1	52.9
GPT-4.1	**59.5**	58.0	66.8	72.0	62.9	48.7	**45.9**	73.4	**59.6**	71.4	68.8	84.1	**63.9**	63.9	60.4	71.2	70.5
GPT-5 mini	58.3	**60.1**	**69.3**	72.9	**64.4**	**52.5**	43.1	**74.9**	57.9	**73.0**	**70.1**	**86.0**	61.7	**67.5**	59.9	**74.3**	**72.5**
o4-mini	57.7	58.4	67.9	**73.6**	61.8	49.8	45.0	73.3	58.0	72.5	68.0	82.3	62.0	65.5	**60.8**	70.9	69.4
COMET																	
Llama 3.2 11B	71.3	60.5	70.1	80.0	79.4	75.7	59.3	72.1	63.7	59.7	67.6	66.0	68.6	74.3	77.0	67.7	73.3
Gemma 3 12B	87.5	79.1	88.4	91.6	88.6	86.8	70.5	89.6	88.7	75.7	85.5	88.1	82.8	93.3	91.6	86.3	89.3
GPT-4.1	**90.6**	**93.2**	**94.5**	**96.0**	**92.4**	90.9	85.9	92.3	91.8	94.9	95.4	95.2	91.1	94.0	93.6	89.5	95.5
GPT-5 mini	89.8	93.0	94.2	95.7	92.2	91.4	**86.2**	92.2	91.7	**95.1**	**95.5**	**95.9**	**92.0**	94.9	93.4	**91.5**	**95.8**
o4-mini	90.0	92.6	94.1	95.7	92.0	**91.6**	85.4	**92.4**	**91.8**	93.6	95.3	95.9	91.6	**94.9**	**93.6**	91.5	95.5
COMET-KIWI																	
Llama 3.2 11B	70.3	63.2	68.5	78.8	73.2	69.1	56.6	67.7	58.5	62.1	64.1	62.5	66.3	70.2	74.0	59.9	70.3
Gemma 3 12B	82.4	74.4	78.7	**87.1**	80.2	76.1	61.8	79.6	74.0	66.0	74.9	73.7	75.5	81.7	81.5	76.5	79.9
GPT-4.1	83.0	79.4	**80.8**	86.9	81.7	77.0	69.6	81.5	75.4	**78.0**	80.5	77.9	78.9	82.2	81.6	75.5	82.9
GPT-5 mini	83.0	**80.4**	80.8	86.9	81.7	77.2	**70.3**	81.3	75.5	77.5	80.8	78.3	**80.0**	82.5	**82.1**	78.9	83.1
o4-mini	**83.2**	80.3	80.5	86.9	**81.7**	**77.2**	70.1	**81.6**	**75.5**	77.6	**80.8**	**78.4**	80.0	**82.6**	82.0	**79.1**	**83.1**
NIST																	
Llama 3.2 11B	1.5	0.6	1.2	1.8	2.7	1.3	0.5	1.3	0.8	0.2	1.9	0.7	1.1	1.5	2.8	1.4	0.7
Gemma 3 12B	6.3	3.2	5.1	5.2	6.5	6.1	2.7	4.3	6.6	3.3	2.8	4.9	4.4	6.5	6.3	4.6	5.5
GPT-4.1	**7.1**	6.2	7.7	6.4	8.3	7.1	**6.0**	7.3	7.7	6.8	7.4	7.8	**6.8**	7.0	7.2	**7.5**	7.0
GPT-5 mini	7.0	**6.3**	**8.0**	6.5	**8.3**	**7.3**	5.8	**7.4**	**7.7**	**6.9**	**7.5**	**8.0**	6.7	**7.2**	7.1	7.4	**7.1**
o4-mini	7.0	6.3	7.8	**6.5**	8.2	7.2	5.8	7.3	7.7	6.8	7.4	7.9	6.7	7.1	**7.2**	7.2	6.9

Table 26: Average scores per lecture subject using various metrics in LECTRANS slide translation task with the slide images provided.